UNIVERSITÉ DE MONTRÉAL

CENTRE DE RECHERCHE SUR LES TRANSPORTS

CASE POSTALE 6128, SUCCURSALE A, MONTRÉAL, QUÉBEC, H3C 3J7

PLANIFICATION ET OPÉRATION D'UN SYSTÈME D'URGENCES SANTÉ : LA PROBLÉMATIQUE

par

Pierre Trudeau
Jean Choquette
Jacques Ferland
Jean-Marc Rousseau

Cette recherche a été rendue possible grâce à une subvention de F.C.A.R. dans le cadre du programme "Actions spontanées".

Cette étude a été publiée grâce à une subvention du fonds F.C.A.R. pour l'aide et le soutien à la recherche.

Université de Montréal
Centre de recherche sur les transports - Publication #546
Novembre 1987

Dépôt légal - 4e trimestre 1987

TABLE DES MATIÈRES

Page

Liste des figures et tableaux iii

Sommaire iv

CHAPITRE 1 : INTRODUCTION 1

CHAPITRE 2 : QUELQUES EXEMPLES DE SYSTÈMES AMBULANCIERS A TRAVERS LE MONDE 3

2.1 En U.R.S.S. 3
2.1.1 Moscou 3
2.1.2 Kiev 4

2.2 Angleterre 6

2.3 Toronto 8

CHAPITRE 3 : PROCESSUS EN JEU 10

3.1 Les sources d'appel 10
3.2 Les ressources potentielles 13

CHAPITRE 4 : QUALITÉ DE SERVICE ET CHRONOLOGIE D'UNE INTERVENTION 18

4.1 La loi sur la qualité des services 18
4.1.1 Au Québec 18
4.1.2 Aux Etats-Unis 19
4.1.3 En Angleterre 20
4.1.4 Allemagne de l'ouest 22
4.2 Ambiguïtés des règles 24
4.3 Chronologie d'une intervention 24

CHAPITRE 5 : LA PROBLÉMATIQUE 30

5.1 Découpage global du territoire 32
5.2 Définition d'objectifs 32
5.3 Prévision et explication de la demande 33
5.4 Horaire du personnel répondant aux appels 34
5.5 Horaire des véhicules 34
5.6 Horaires des ambulanciers, médecins 36
5.7 Minimisation du délai interne 36
5.8 Minimisation du délai externe 39
5.9 Tactiques opérationnelles 42
5.9.1 Introduction 42
5.9.2 Enumération des stratégies de localisation/relocalisation 45

TABLE DES MATIÈRES

	Page
CHAPITRE 6 : RECHERCHE BIBLIOGRAPHIQUE ET SOLUTIONS RETENUES	50
6.1 Introduction	50
6.2 Découpage global du territoire	50
6.3 Définition d'objectifs	50
6.4 Prévision et explication de la demande	51
6.5 Horaire du personnel répondant aux appels	52
6.6 Horaire des véhicules	52
6.7 Horaires des ambulanciers	52
6.8 Minimisation des délais interne et externe	53
6.9 Tactiques opérationnelles	53
6.9.1 Critères de sélection d'une approche de solution	54
6.9.2 Sans relocalisation	55
6.9.2.1 Problèmes de recouvrement	55
6.9.2.2 Modèles stochastiques	57
6.9.2.3 Approche hybride	68
6.9.3 Avec relocalisations	68
6.9.4 Nos réalisations	70
6.9.4.1 Sans relocalisation	71
6.9.4.2 Avec relocalisation	74
Références	76

LISTE DES FIGURES ET TABLEAUX

Page

Tableau 1 : Standards de performance à atteindre en Angleterre 23

Figure 1 : Chronologie d'une intervention 25

Figure 2 : Problématique d'un système ambulancier 31

Figure 3 : Demande horaire de véhicules 35

Figure 4 : Relocalisation avec postes d'attente fixes 47

SOMMAIRE

La majorité des centres urbains sont aujourd'hui dotés de systèmes de véhicules d'urgence (le service ambulancier, la police, les pompiers). L'exploitation de tels systèmes suscite des problèmes considérables, notamment en raison de la taille des problèmes et de la nature aléatoire de plusieurs des processus en jeu.

Bien que les outils développés puissent facilement être adaptés au contexte de la police ou du service des incendies, nous nous sommes concentrés dans cette étude sur le système ambulancier. Tout au long du projet nous avons bénéficié de l'appui et de la disponibilité d'Urgences santé grâce à qui nous avons approfondi notre connaissance de la problématique et obtenu d'indispensables données sur leurs opérations.

Nous avons orienté nos recherches dans trois directions. Il y a d'abord l'axe qui conduit à la confection des horaires des ambulanciers. Il comprend un modèle d'explication de la demande quotidienne et un modèle de prévision horaire de la demande. Une fois la demande prédite on peut construire l'horaire des véhicules puis celui des ambulanciers de manière à avoir en tout temps un nombre adéquat de ressources.

Vient ensuite tout le champs des stratégies opérationnelles qui visent à obtenir les meilleures performances possibles avec les ressources mises en service. Nous insistons surtout sur les aspects localisation et relocalisation dynamique des véhicules sur le territoire. En ce qui a trait à la localisation, quelques modèles analytiques permettant dans certains contextes d'identifier des configurations près de l'optimum ont été implantés.

Pour permettre enfin l'évaluation de stratégies dans un contexte le plus réaliste possible, nous avons jugé essentiel de nous doter d'un logiciel de simulation. Dans un contexte où des vies humaines sont en jeu, on peut donc sans risque mettre en évidence les aspects positifs et négatifs d'une nouvelle stratégie avant de l'implanter réellement.

CHAPITRE 1
INTRODUCTION

De nos jours, la plupart des centres urbains sont dotés d'un système ambulancier ayant pour mandat de répondre aux urgences médicales. L'exploitation d'un tel système suscite des problèmes considérables. D'autres systèmes évoluent dans des contextes tout aussi complexes, mais la qualité du service a ici des répercussions directes sur la santé d'une population et s'exprime, entre autres, en termes de vies humaines sauvées.

En juin 1984, le C.R.T. obtenait une subvention auprès du fonds F.C.A.C. dans la catégorie "Actions spontanées". Le projet avait pour titre "Planification et exploitation d'un système d'urgences santé". Il avait pour objectif de "développer un ensemble cohérent et complet d'outils de planification et de gestion pour l'allocation et la répartition des véhicules d'urgence". Au terme du projet, le présent document se veut à la fois un exposé de la problématique d'un système ambulancier en même temps qu'une synthèse de nos travaux. Nous y énumérons donc dans quel cadre (contraintes) un système ambulancier évolue et quelles formes peuvent prendre les solutions en vue d'atteindre les objectifs choisis.

Avant d'aller plus loin, nous tenons à souligner que tout au long du projet nous avons pu compter sur l'étroite collaboration d'Urgences santé, l'organisme chargé de l'exploitation du système ambulancier à Montréal et dans les environs. Un merci tout particulier à M. Roger Gauthier, conseiller en recherche opérationnelle, ainsi qu'à tous les autres qui ont fait preuve de disponibilité et maintenu leur intérêt et leur soutien au projet.

Toutes les données utilisées au cours du projet furent fournies par Urgences santé et reflètent donc leur situation. Nous ne nous sommes cependant pas limités au seul cas de Montréal. Nous avons voulu, au contraire, étudier la plus grande diversité possible pour mieux situer le cas particulier de Montréal et les types de solution envisageables dans un tel con-

texte. Notre exposé ne se veut pas exhaustif et, en particulier, les décisions d'ordre médical n'étant pas de notre ressort elles seront peu abordées. Dans ce document, aucun résultat n'est présenté comme tel. Au fil de l'exposé, toutefois, nous référons le lecteur aux publications pertinentes issues du projet, en les situant par rapport à notre démarche et par rapport aux alternatives possibles.

Dans le chapitre 2 nous examinons brièvement quelques systèmes ambulanciers. Au chapitre 3 nous décrivons les divers processus en jeu, i.e. ce qui constitue la matière première du système. Les critères de qualité de service et la chronologie d'une intervention sont abordés au chapitre 4. Puis, au chapitre 5 nous dressons une liste des divers problèmes qui surgissent. Au chapitre 6, enfin, nous exposons et motivons des solutions que nous avons retenues pour les problèmes que nous avons choisi d'attaquer.

Le but de ce document est de sensibiliser le lecteur :

- au fonctionnement d'un système ambulancier;
- aux problèmes qu'il suscite;
- aux solutions possibles (proposées dans la littérature ou par nous);
- au travail qu'il reste à faire.

CHAPITRE 2

QUELQUES EXEMPLES DE SYSTÈMES AMBULANCIERS À TRAVERS LE MONDE

Dans ce qui suit, nous retrouvons une description de quelques systèmes ambulanciers répartis à travers le monde. Nous rendons compte de leur situation à l'époque où les références ont été publiées. Quant à Montréal, son organisation sera décrite en détail au cours des chapitres subséquents.

2.1 En U.R.S.S. [U.K., 1976]

On retrouve en U.R.S.S. un organisme qui a la charge de toute la recherche et l'administration des services médicaux d'urgence. Il s'agit de l'institut Sklifosovsky situé à Moscou et qui est en même temps un centre hospitalier. L'institut reçoit des rapports en provenance de toutes les parties de l'U.R.S.S. qu'il analyse par type de maladie ou d'accident. Un rapport est rédigé par type de maladie/accident et il est entre autre soumis aux sociétés spécialisées dans les différents types de maladies/accidents, donnant lieu ainsi à diverses recommendations pour un meilleur traitement de la part des services médicaux d'urgence.

2.1.1 Moscou

Pour les fins du service ambulancier, la ville est subdivisée en huit zones (une centrale et sept périphériques) de population à peu près égale. Cette subdivision a été faite en tenant compte des mouvements de la population au cours de la journée de telle sorte que la majorité de la population habitant une zone y travaille également. Dans chaque zone on retrouve un hôpital pour soins d'urgence et une caserne d'ambulances adjacente qui assure le contrôle de la zone. Il y a toutefois un contrôle central qui est situé à l'institut Sklifosovsky dans la zone centrale. Il y a en plus trois ou quatre autres casernes par zone pour un total de trente-quatre pour toute la ville. Dans chaque zone d'autres hôpitaux ont également à tour de rôle des lits réservés pour l'urgence. Au total 125 000

lits sont réservés pour l'urgence et sont sous le contrôle du service d'urgence médicale.

Tout citoyen peut recevoir des soins en cas d'urgence en composant un seul numéro différent de celui pour faire appel à la police ou aux pompiers. Les appels sont logés au contrôle central qui les achemine au sous-contrôle de la zone appropriée. Dans les vingt-quatre heures d'une journée la période de pointe s'étend de 18h00 à 24h00. Il y a jusqu'à 5 000 appels par jour et l'hiver est la saison de pointe en raison des surfaces glissantes. Selon le moment du jour, de 200 à 500 équipes sont en service.

Deux fois par jour le contrôle central et chacun des sous-contrôles reçoivent l'information sur l'état des lits d'urgence. Le contrôle central a le pouvoir de décider ultimement à quel hôpital le patient sera amené. L'établissement de temps de réponse à atteindre est la responsabilité de l'institut. Nous ignorons les objectifs qu'ils se sont fixés.

Les huit zones ont des caractéristiques démographiques différentes (par exemple, certaines ont une majorité de personnes âgées) et les équipes d'ambulanciers sont affectées en conséquence (par exemple, des spécialistes pour le coeur dans les zones de personnes âgées).

Le service d'urgence est également responsable du transport des patients contagieux et des cadavres.

2.1.2 Kiev

Pour une population totale de 2.1 millions d'habitants, on y retrouve 125 ambulances (dont 17 ambulances spécialisées), 400 médecins et 650 auxilières. La ville est divisée en 10 districts chacun ayant ses casernes. On porte beaucoup d'attention à la coopération entre le service ambulancier et les hôpitaux d'une part, et entre les médecins du service ambulancier et ceux des polycliniques d'autre part.

En général les casernes sont grandes, pouvant abriter de 8 à 21 véhicules et équipages.

En 1974, sur les 550 000 appels reçus, la répartition s'établissait comme suit :

cas cardiaques	72%
trauma	10%
patient ne nécessitant qu'un transport	10%
maternité	4%
faux numéros	4%

et 60% d'entre eux furent traités par les ambulances non spécialisées.

On fait continuellement l'étude du diagnostic initial versus le diagnostic final.

En tout temps, huit proposés ayant une formation médicale recueillent les appels sous la supervision d'un médecin. Les fiches d'appel sont ensuite transmises mécaniquement à une équipe de trois répartiteurs supervisés par un médecin. Ils transmettent l'appel au sous-contrôle de la zone appropriée.

Les soins sont offerts en cas de :

- accidents avec blessures sérieuses, traumas, électrocution, brûlures, engelure, suffocation, intoxication, tentative de suicide;
- malaise soudain mettant la vie en danger;
- désordre mental aigu accompagné de gestes d'agression;
- accouchement.

Le service ambulancier ne se déplace pas en cas de :

- malaise ne mettant pas la vie en danger;
- blessure mineure;
- lorsque le patient est sous l'effet de l'alcool.

2.2 Angleterre [Orcon, 1974], [Groom, 1975], [Holloway, 1976b], [Raitt, 1976a], [Raitt, 1976b], [Raitt, 1977]

En avril 1974, la responsabilité du service ambulancier en Angleterre passa des autorités locales au N.H.S. (National Health Service). De quarante-cinq services au niveau des comtés et soixante-huit au niveau des villes, on passa à quarante-cinq services, urbains (M.H.A. - Metropolitan Health Authority) ou plus ruraux (A.H.A. - Area Health Authority). Tous les services sont sous le contrôle du N.H.S. Chacun fonctionne cependant de manière indépendante sur son territoire et il n'y a pas vraiment de coopération entre les divers services. Jusqu'à cette réorganisation, aucun standard national n'existait pour le service aux patients. Les autorités locales pouvaient s'en fixer de manière indépendante. On profita de la réorganisation pour commander une étude sur le sujet. Les grandes lignes du rapport final sont présentées à la section 4.1.3.

Les appels d'urgence ne représentent que 6% de tous les patients transportés. L'urgence résulte habituellement d'un accident ou d'un malaise soudain. Le reste des appels, ceux dits non-urgents, correspondent aux appels inter-établissement de Montréal. Il faut dire qu'en Angleterre le service ambulancier a le devoir de transporter un patient qui ne peut se rendre par transport public au lieu d'un traitement médical. Ces cas augmentent d'ailleurs constamment à cause de la disparition de certaines liaisons par autobus en milieu urbain. On considère toutefois que les procédures d'autorisation de ces transports ne sont pas assez strictes.

Pour répondre aux appels (urgents et non-urgents) on semble considéré deux types d'organisation de la flotte :

i) une flotte unique qui répond aux appels urgents comme aux appels non-urgents avec un nombre prédéfini réservé à l'urgence. Ainsi lorsqu'un véhicule d'urgence devient occupé c'est le prochain véhicule disponible (qu'il soit affecté à l'urgence ou non) qui le remplace. En général, lorsqu'une urgence survient, c'est le véhicule disponible (qu'il soit

affecté à l'urgence ou non) le plus près qui y répond;

ii) deux flottes indépendantes : une pour les appels urgents et une pour les appels non-urgents.

Il arrive que l'on utilise une forme limitée de relocalisation où on tente de toujours couvrir les casernes les plus occupées. On souligne que cela est surtout utile en milieu urbain où la distance entre deux casernes est relativement courte. On prévoit également dans certains cas (en périodes de pointe) localiser des véhicules en attente en d'autres lieux que les casernes, en plus de celles-ci, pour avoir une meilleure couverture.

On observe qu'en général, pour un nombre égal de véhicules, l'utilisation d'une flotte unique est nettement plus efficace que le système à deux flottes. Il est courant qu'en milieu rural la couverture soit partiellement assurée par des ambulanciers en disponibilité à domicile. Lorsqu'appelé, l'ambulancier doit d'abord se rendre à la caserne où une ambulance l'attend. Cela peut augmenter le temps de réponse jusqu'à sept minutes. Cette pratique est jugée peu souhaitable.

En comté rural, l'organisation est souvent la suivante :

- le jour (08h00-16h00, ≈50% des appels) : une flotte unique
- le soir (16h00-24h00, ≈30% des appels) : des flottes séparées mais avec relocalisation limitée
- la nuit (00h00-08h00, ≈20% des appels) : une flotte unique (il n'y a pas d'appels dits non-urgents la nuit) sans relocalisation et avec certains ambulanciers à domicile sur appel.

On a souvent observé en Angleterre que le service pour les appels non-urgents était insatisfaisant. La raison principale en est la piètre collaboration entre le service ambulancier et les hôpitaux. Un service où les hôpitaux déterminent les heures de rendez-vous sans consulter le service ambulancier génère un gaspillage de ressources en plus d'occasionner

des inconvénients aux patients. Il est donc primordial que les deux services collaborent étroitement. À cet effet, on suggère dans certains cas (dans les C.H. les plus achalandés) la présence d'agents de liaison. Ils auraient la charge de coordonner les activités C.H./service ambulancier, donner des heures de rendez-vous réalistes, répartir la demande, regrouper des demandes, assister les ambulanciers au besoin, tenir les patients au courant de tout délai dans les transports pour réduire leur anxiété, ... On souligne que les tournées pour les appels non-urgents devraient être planifiées à l'avance (l'après-midi précédent par exemple) et qu'on devrait pouvoir refuser, sauf exception, les demandes reçues après un certain délai limite.

L'utilisation d'un service de voitures pourrait également grandement améliorer la qualité du service ambulancier en libérant les ambulanciers pour leur permettre de mieux s'occuper des patients qui ont de vrais besoins médicaux, et ce tout en réduisant le coût total du service ambulancier.

2.3 Toronto [Allaire, 1986], [Toronto, 1986]

À Toronto (Department of Ambulance Services of the Municipality of Metropolitan Toronto) le service ambulancier dessert une population de 2.5 millions de résidents. Durant les heures d'affaire la population atteint les 4 millions. Le territoire desservi est de 632 km^2 et 308 000 appels ont été reçus en 1985.

En décembre 1975, le service ambulancier fut consolidé. Un regroupement de six municipalités en a maintenant la charge : Toronto, York, Etobicoke, Scarborough, North York et East York.

La répartition des appels en 1985 était la suivante :

Appels d'urgence (16% d'entre eux par des tiers: police pompiers)	54%
Appels non-urgents	31%
Événements spéciaux	9%
Autres appels	6%

pour un temps moyen de réponse (i.e. de l'affectation jusqu'à l'arrivée sur les lieux de l'appel) aux appels urgents de 6,21 minutes.

Toronto dispose de 40 casernes disséminées sur son territoire. Les types de service offerts sont : "Basic Life Support", "Advanced Life Support" et le transport de patients. Environ 40% des appels sont de type "interétablissements" dont 15% sont effectués par des MPU (Multi-Patient-Units) qui sont des autobus de cinq civiè- res et six places. Ces véhicules font en moyenne vingt transports par jour par véhicule. Trois employés sont à bord et ils effectuent des transports planifiés à l'avance. Les gros centres hospitaliers bénéficient d'agents de liaison pour coordonner le service avec les ambulanciers.

Le territoire est divisé en quatre secteurs (chacun regroupant plusieurs casernes) opérés de manière indépendante, i.e. sans coopération en cas de manque d'effectif. Il n'y a pas non plus de relocalisation des véhicules. La prise d'appel est faite par un personnel n'ayant pas de formation en sciences infirmières mais seulement une formation restreinte.

Après ce bref survol de quelques systèmes ambulanciers, nous en abordons au cours des trois prochains chapitres la description plus formelle. Nous commençons par l'examen des divers processus en jeu.

CHAPITRE 3

PROCESSUS EN JEU

Dans ce chapitre, nous présentons les protagonistes d'un système ambulancier. Il y a d'abord ceux que l'on subit, i.e. sur lesquels on n'a aucun contrôle : les sources d'appel. Puis il y a ceux dont on se dote pour répondre aux premiers : les ressources.

3.1 Les sources d'appel

Lorsqu'on entend parler du service ambulancier c'est la plupart du temps par la voie des médias ou lorsqu'un proche a dû y recourir. On en a donc une vision bien limitée et on pourrait penser que ces services ne sont sollicités qu'en cas d'extrême urgence, lorsqu'il y a explosion, incendie, accident de la route, crise cardiaque, etc. On a donc tendance à faire le rapprochement entre "ambulance" et "vie en danger". Or rien n'est généralement plus faux. Généralement parce que c'est, là aussi, une question de choix. L'organisme en charge du système peut choisir de ne répondre qu'aux appels d'extrême urgence comme il peut choisir d'offrir une gamme beaucoup plus large de services.

À Montréal, Urgences santé offre un éventail de services des plus étendu. Voici brièvement le genre d'appel qu'il peut recevoir :

i) Appel ne donnant lieu à aucun transport :

- demande de renseignement : un citoyen cherche, par exemple, le numéro de téléphone d'un dentiste ou recherche la salle d'urgence la plus près de chez lui et fait appel à Urgences santé pour l'aider dans cette tâche. Il s'agit là d'un cas d'ignorance ou de mauvaise foi puisqu'Urgences santé n'a pas le mandat de répondre à ce genre d'appel.
- abandon de la demande : après discussion avec Urgences santé, le demandeur décide de ne pas faire appel à ses services pour des raisons de coût, délai ou simplement parce que la réceptionniste a pu répondre à son besoin sans qu'un transport soit nécessaire.

ii) Appels dits d'urgence :

On pourrait emprunter ici la définition d'urgence utilisée à Toronto [Ontario, 1970], que l'on peut traduire ainsi : "situation où un délai dans la réponse à un appel peut mettre en danger la vie, un membre ou un organe vital d'un patient". Cette définition correspond à la catégorie des appels nécessitant une intervention la plus rapide possible

- arrêt cardio-respiratoire
- accident de la route
- incendie, explosion
- polytraumatisé.

Mais il y a également ceux nécessitant une intervention mais pas nécessairement immédiate

- maux de ventre
- étourdissement
- cheville foulée.

Il est à noter que dans tous ces cas l'appel ne donne pas nécessairement lieu à un transport du patient vers un centre hospitalier (C.H.). Il suffit parfois de traiter le patient sur place.

iii) Événements spéciaux :

À l'occasion d'événements spéciaux il arrive que l'on demande de dépêcher une ou plusieurs ressources qui seront rattachées exclusivement à ces événements. Parmi ceux-ci on compte le Grand Prix automobile de Formule 1, les concerts rock, le parc d'amusement de La Ronde les soirs d'été et les défilés dans les rues.

iv) Appels dits de transport inter-établissement :

Ces appels forment véritablement une classe à part et sont souvent traités comme tel alors qu'un sous-ensemble des ressources leur est dédié (c'était le cas auparavant chez Urgences santé). Ces cas sont rarement de première urgence. Ils sont la plupart du temps reliés à des tests que doivent subir un patient dans un C.H. Ils ont aussi la particularité d'être connus, en partie, à l'avance (un jour) de sorte

qu'on peut s'y préparer. On y retrouve :

- le transport d'un centre hospitalier à un autre centre hospitalier, le plus souvent pour permettre au patient de subir un test qui n'est pas offert au premier centre hospitalier;
- le transport d'un centre hospitalier à un domicile ou d'un domicile à un centre hospitalier : le patient ne peut pas utiliser les transports publics pour revenir ou aller subir un test médical. On compte également une partie de la population qui bénéficie de ce service gratuitement (personnes âgées, assistés sociaux).

Comme on le voit, les appels sont d'une grande diversité. Il n'est donc pas surprenant que certains organismes leur associent des priorités diverses pour apporter une réponse appropriée à chacun d'eux. En pratique, la priorité décrit la rapidité d'intervention requise, ou plus précisément, "un estimé conservateur de la rapidité d'intervention requise" puisque l'état du patient ne peut être évalué que via le téléphone. Or le plus souvent le demandeur n'a pas de connaissances médicales particulières, et en cas de doute, on préfèrera attacher un niveau de priorité supérieur.

Chez Urgences santé, le personnel qui répond aux appels a des connaissances médicales et peut (et dans la pratique le fait) associer une priorité à chaque appel. Il y a trois niveaux de priorité, à la fois pour les appels dits d'urgence (les niveaux STAT, 2 et 4) et ceux d'interétablissement (les niveaux 1, 3 et 5). À chaque niveau correspond un temps de réponse qui se veut une borne supérieure à toute intervention de ce type. (Chez Urgences santé, le temps de réponse est défini comme étant la différence entre le moment de l'affectation de l'appel et le moment d'arrivée sur les lieux de l'appel d'une ressource.)

Priorité	Borne supérieure sur le temps de réponse
STAT, 1	5 min
2, 3	20 min
4, 5	60 min

L'affectation d'une priorité a en général deux conséquences :

i) tout d'abord elle influence l'ordre de traitement des appels. Parmi les appels reçus et en attente de traitement par le répartiteur celui de priorité plus élevée passera toujours devant celui de moindre priorité et ce peu importe l'heure de réception de chacun des deux appels. Globalement, c'est donc un système de type

   - "premier arrivé premier servi" à l'intérieur de la classe d'appels de même priorité;
   - préséance absolue d'une classe de priorité plus élevée sur toutes celles de priorité plus basse.

ii) de plus, la ou les ressources affectées pour répondre à un appel pourront ou non se déplacer à grande vitesse, utiliser gyrophares et sirènes et passer sur les feux rouges selon le niveau de priorité. Dans la négative les véhicules dépêchés respectent le code de la route comme tout autre véhicule.

3.2 Les ressources potentielles

Deux types de ressources sont impliqués dans le processus d'intervention :

- les ressources humaines
- les ressources physiques (véhicules, équipement, outils de communications,...).

Nous ne traitons que des premières, et parmi celles-ci nous nous limitons à celles qui ont un contact avec le patient. D'autres intervenants, comme les répartiteurs par exemple, ont un rôle important à jouer et feront l'objet d'une discussion à la section 4.2.

Les ressources humaines se distinguent par le niveau de leur formation ce qui détermine leur degré d'intervention. Elles sont présentes à trois moments critiques différents :

- lors de la réception de l'appel
- lors de l'arrivée sur les lieux de l'appel
- lors de l'arrivée au centre hospitalier.

À chacun de ces moments, la formation du personnel joue un rôle primordial sur la suite des événements. De fait, le système d'urgence doit prendre des décisions quant à la formation qu'il exige de chacun des intervenants.

a) réception de l'appel :

Dès cette première prise de contact, une opération de filtrage peut ou non être effectuée. Considérons d'abord la situation où le répondant n'a aucune formation médicale. Il ne peut associer aucune priorité aux appels, ni en éliminer. Par conséquent, tous les appels ont la même priorité et l'action prise est uniforme : tous les véhicules affectés aux appels circulent à la même vitesse (probablement élevée dans ce cas). En revanche, le seul renseignement requis est l'adresse ce qui a pour effet d'accélérer l'intervention de sorte qu'en principe une ressource sera affectée plus rapidement.

Dans le cas où le répondant a certaines qualifications médicales (infirmièr(e), médecin) on peut lui permettre de filtrer les appels et d'attribuer des priorités. C'est le cas chez Urgences santé où tous les répondants sont des infirmièr(e)s et où en tout temps un médecin est présent comme personne ressource. Le répondant doit par contre recueillir plus d'information que dans le premier cas : non seulement l'adresse du demandeur mais aussi une description de son état. Cela prend évidemment plus de temps mais permet, par le filtrage, de réduire le nombre de transports nécessaires. Au moment où le répartiteur reçoit la fiche appel il y a ainsi plus de chance qu'une ressource soit disponible.

b) arrivée sur les lieux de l'appel :

C'est à ce moment qu'a lieu le premier contact physique avec le patient.

- Si l'intervenant n'a qu'une formation médicale superficielle il n'agit que comme transporteur vers un CH où le patient sera vraiment pris en charge.

- Si l'intervenant a une formation restreinte son rôle se "limite" à stabiliser l'état du patient sans toutefois pouvoir poser certains actes. Tout comme ci-dessus, la première véritable intervention médicale se fera au CH.
- Si l'intervenant est paramédic ou médecin, il stabilise l'état du patient sur place et peut même compléter certains traitements avant de le confier au CH si une hospitalisation s'avère nécessaire.

c) arrivée au CH :

Une fois au CH le patient y est pris en charge par le personnel des lieux . Ce personnel n'est en général pas sous l'autorité du système ambulancier qui n'a donc aucun droit de regard sur sa formation. Malgré tout, le système ambulancier a certaines responsabilités. Il doit s'assurer que le CH choisi est approprié, i.e. qu'il a les spécialistes nécessaires (s'il y a lieu) et que son urgence n'a pas atteint un niveau d'occupation nécessitant un détournement vers un autre CH.

Le fait que le premier intervenant puisse ou non poser des actes médicaux influence la qualité du service offert. Plus le personnel est qualifié, cependant, plus il coûte cher. Pour un budget fixe donc, on peut se demander laquelle des options suivantes est préférable :

- ne disposer que de premiers intervenants qualifiés mais en petit nombre : étant peu nombreux ils ont en moyenne de plus grandes distances à parcourir et arrivent plus tard sur les lieux de l'appel, mais alors ils prennent contrôle du patient;
- ne disposer que de premiers intervenants moins qualifiés mais en plus grand nombre : ils arrivent plus rapidement sur les lieux de l'appel mais le patient n'est vraiment pris en charge que plus tard au CH;
- une combinaison des deux premières options où la formation des premiers intervenants est variable.

C'est la troisième option qui a été retenue chez Urgences santé. Mais cette troisième option permet de nombreuses combinaisons. Plus spéci-

fiquement chez Urgences santé on retrouve :

- une flotte d'ambulances qui assurent le transport des patients (Chaque véhicule compte deux ambulanciers qui ont une formation restreinte et se "limitent" à prodiguer les premiers soins);
- une flotte de voitures comptant un chauffeur et un médecin (Ces voitures ne peuvent pas effectuer le transport de patients);
- quelques superviseurs qui ont une formation de paramédic et qui conduisent des véhicules tout-terrain (un superviseur par véhicule). Ils ne peuvent faire de transport de patient et chacun a la responsabilité d'un territoire (trois au total). Ils en font la patrouille et assistent au besoin les autres ressources. Ce sont eux qui supervisent les opérations en cas d'incident majeur nécessitant l'intervention de plusieurs ressources.

Suite à un appel, l'infirmièr(e) répondant décide du nombre et des types de ressource à envoyer. Il peut s'agir d'ambulanciers seulement (1 ou plusieurs véhicules) de médecin(s) seulement (1 ou plusieurs véhicules) ou d'ambulanciers et de médecins.

À simple titre de comparaison, citons le cas de Moscou [U.K., 1976] où en réponse à un appel d'urgence l'une ou l'autre des trois catégories d'ambulance suivantes peut être dépêchée sur les lieux :

i) ambulance générale : c'est la plus courante; elle s'occupe de la majorité des cas de trauma et de maladie. On y retrouve un médecin, un auxiliaire, un(e) infirmier(e) et un chauffeur.

ii) ambulance spécialisée : il n'y en a qu'un petit nombre. Chacune a un personnel et un équipement spécialisé (par exemple, neurologique). On y retrouve un médecin, deux auxiliaires (dont un technicien de laboratoire), un(e) infirmier(e) et un chauffeur.

iii) ambulance sans médecin: On y retrouve que des auxiliaires qui s'occupent principalement des patients externes des polycliniques et du transport interétablissement.

La qualité de son service est évidemment une préoccupation majeure de tout système ambulancier. La définition d'objectifs à cet égard fait l'objet du chapitre suivant. On y trouve prétexte pour fournir une description chronologique détaillée du déroulement d'une intervention suite à un appel.

CHAPITRE 4

QUALITÉ DE SERVICE ET CHRONOLOGIE D'UNE INTERVENTION

Le service ambulancier est onéreux. On estimait en 1986 qu'à Montréal, mettre une ambulance en service 24 heures par jour, tous les jours de l'année coûtait environ 500 000 $CDN! Il est donc primordial d'utiliser les ressources disponibles de la meilleure façon possible, car les économies sont appréciables. À qualité de service égale, l'utilisation d'un véhicule en moins représente un gain non négligeable.

La note finale dépend, en gros, de deux facteurs :

i) les objectifs de qualité de service que l'on aura fixés a priori;

ii) la stratégie utilisée pour atteindre ces objectifs.

Plus lesdits objectifs sont exigeants, plus il en coûte cher de les réaliser. Il est donc légitime de recourir à des outils tactiques de pointe pour planifier adéquatement et réduire ainsi le volume de ressources requises. Nous abordons maintenant la forme que peuvent prendre ces objectifs. Les stratégies possibles sont globalement traitées au chapitre 5.

4.1 La loi sur la qualité des services

Le système ambulancier touchant à la santé du public il n'est pas surprenant que les autorités publiques aient le plus souvent édicté des standards à respecter.

4.1.1 Au Québec

Au Québec [Québec, 1983] dans la loi sur la protection de la santé publique, on retrouve le paragraphe suivant :

"1. Détermination des zones de services d'ambulance

Une zone de services d'ambulance est un territoire qui est délimité par une ou plusieurs municipalités d'une même région socio-sanitaire.

Les points de service d'une telle zone sont :

- le centre démographique;
- le carrefour des routes importantes;
- le centre géographique;
- le centre socio-économique,

duquel toutes les parties du territoire municipalisé peuvent être rejointes par une ambulance dans un délai maximum de 30 minutes à compter du moment de l'appel, dans des conditions normales de température, de circulation et d'état du réseau routier relié directement au réseau provincisé."

À Montréal, les objectifs fixés par Urgences santé ont déjà été mentionnés à la section 3.1 (p. 11).

4.1.2 Aux Etats-Unis

En 1973, les américains mettaient en vigueur le "EMS Act" (Emergency Medical Services) [U.S., 1974]. En plus de prescrire ce que doit être un service d'urgences médicales cette loi comporte des règlements sur certaines performances à atteindre. On y demande que

- en milieu urbain, 95% de la demande soit couvert en 10 minutes;
- en milieu rural, 95% de la demande soit couvert en 30 minutes.

À titre d'exemple, mentionnons qu'à Seattle [Mayer, 1981] le système comporte trois niveaux. Dans le cas d'un appel d'urgence (où la vie du patient peut être menacée) on affecte un premier intervenant qui est l'unité de la caserne de pompiers la plus près de l'incident. Ce sera alors ou bien une "unité d'aide" ou un camion de pompier. Ces intervenants prodiguent les soins de base en attendant l'arrivée de l'un des quatre véhicules de paramédic.

Sur les 20 000 appels reçus en 1976 un peu moins de 50% a nécessité la présence d'un paramédic. Dans environ 95% des appels l'unité de la caserne était sur les lieux en moins de 6 minutes. Dans plus de 95% des appels nécessitant un paramédic, ce dernier était sur les lieux en moins de 12 minutes. Le temps moyen de réponse du premier intervenant était environ

trois minutes comparé à un peu moins de six minutes pour les paramédics.

À Austin [Eaton, 1985] en 1975 les autorités se fixèrent comme objectif que tous les appels soient couverts en moins de 5 minutes. Il semble que la notion de couverture doive être interprétée ici en supposant que chaque poste d'attente est occupé en permanence.

4.1.3 En Angleterre

Le D.H.S.S. (Department of Health and Social Security) profita de la réorganisation nationale des services ambulanciers (voir chapitre 2) pour commander une étude sur les standards de qualité.

En juillet 1974, le rapport final fut présenté [Orcon, 1974]. On avait étudié 12 services considérés représentatifs. Pour les besoins de l'étude on récueillit spécialement un mois de données pour les appels d'urgence et une semaine de données pour les appels non-urgents. On y considère que la qualité du service pour les urgences peut être évaluée à l'aide des deux statistiques suivantes :

i) le temps de mise en oeuvre : c'est le temps qui s'écoule entre la réception de l'appel (le début de l'appel) et l'affectation du véhicule approprié

ii) le temps total de réponse : c'est le temps qui s'écoule entre la réception de l'appel et l'arrivée sur les lieux de la ressource affectée. C'est ce temps qui est de première importance pour la qualité du service.

On considère également des appels semi-urgents qui ne requièrent pas l'arrivée immédiate d'une ressource mais pour lesquels l'arrivée à l'hôpital du patient doit se faire à l'intérieur d'un certain délai (souvent une heure).

Finalement, les appels dits non-urgents sont ceux impliquant le

transport de patients pour des traitements à l'hôpital. Ces transports se font du (à l'aller) ou vers (au retour) le domicile du patient et sont en général requis lorsque le patient n'a pas d'autre moyen pour se rendre à l'hôpital (par exemple en région rurale lorsqu'il n'y a pas de service d'autobus intercité). Cette demande est le plus souvent connue à l'avance et peut donc être planifiée. Ce service représente 94% des patients transportés.

Même si la priorité est accordée aux appels urgents et semi-urgents, on estime que l'on doit se préoccuper de la qualité du service pour les appels non-urgents également. Le service idéal pour un appel non-urgent serait le suivant : arrivée juste pour l'heure du rendez-vous par un trajet direct, traitement, puis départ immédiat et retour à domicile sans détour. On a donc définit cinq variables pour identifier dans quelle mesure on s'éloigne du service idéal :

i) temps additionnel par rapport à un trajet direct pour se rendre au centre de traitement;

ii) temps d'attente avant le rendez-vous;

iii) temps en retard pour le rendez-vous;

iv) temps d'attente après le traitement avant que le trajet de retour ne s'amorce;

v) temps additionnel par rapport à un trajet direct pour revenir au centre de traitement.

Comme les items i) et v) sont difficiles à mesurer on les a abandonnés au profit de la charge de l'ambulance (exprimé en nombre de patients). Enfin, puisque certains patients ont des rendez-vous pour lesquels il est plus important qu'ils soient à l'heure, on a défini des objectifs différents pour eux.

À partir de ces observations on en arriva à formuler des standards de qualité à atteindre. On mentionne qu'idéalement ces standards devraient refléter les performances que l'on désire atteindre. On ajoute toutefois que dans le contexte actuel (étant donné les ressources disponibles et les

façons de procéder) de telles performances sont hors de portée. On a plutôt défini des standards qui pouvaient être atteints par la majorité des services ambulanciers. Ces standards apparaissent dans le tableau 1.

La question de savoir si ces objectifs devraient être atteints à chaque période du jour (jour, soir et nuit par exemple), dans chaque région, ou si une moyenne pondérée devrait être utilisée relève des autorités.

D'autre part, l'expérience a montré la difficulté d'obtenir des données sur les performances pour les appels non-urgents.

On souligne en terminant que le plus long temps de réponse (calculé comme le plus grand temps de mise en oeuvre et le plus long déplacement, en supposant toutes les ressources disponibles), est un critère utile. Il assure une qualité minimale de service pour tout le territoire.

4.1.4 Allemagne de l'ouest

Durant les années 70, la République Fédérale d'Allemagne a voté des lois et des règlements sur les services d'urgences. En voici quelques éléments [Bach, 1984] :

- la responsabilité de ces services incombe aux comtés, aux municipalités ou à des organisations régionales;
- chaque système doit assurer la couverture du territoire dont il a la charge;
- le critère de qualité est le temps de réponse (la définition précise n'est pas donnée) et il faut qu'en tout point du territoire ce temps soit en deça d'un temps maximal. Dans au moins un cas (la ville de Dortmund, 620 000 habitants en 1979, 108 milles^2), ce temps maximal est de 8 minutes;
- la tâche des autorités est de minimiser le nombre de casernes tout en satisfaisant la contrainte du temps maximal de réponse.

Tableau 1 : Standards de performance à atteindre en Angleterre

Mesures de performance	Standards à atteindre		
	%	zone urbaine	zone rurale
Appels urgents			
temps de mise en oeuvre	95	3 minutes	3 minutes
temps total de réponse	50	7 minutes	8 minutes
	95	14 minutes	20 minutes
Appels semi-urgents			
temps d'arrivée par rapport au temps prévu d'arrivée	95	5 minutes en retard	5 minutes en retard
Appels non-urgents			
charge de l'ambulance à l'aller et au retour du centre de traitement	50	2 patients	2 patients
	95	8 patients	8 patients
Temps d'arrivée par rapport au temps prévu			
a) cas général	5	40 min en avance	60 min en avance
	25	20 min en avance	30 min en avance
	75	15 min en retard	20 min en retard
	95	40 min en retard	60 min en retard
b) cas spécial	5	40 min en avance	60 min en avance
	25	20 min en avance	30 min en avance
	75	10 min en retard	15 min en retard
	95	20 min en retard	30 min en retard
Attente après le traitement	50	30 minutes	30 minutes
	95	60 minutes	60 minutes

4.2 Ambiguïtés des règles

À première vue, les critères énoncés ci-dessus semblent clairement définis. Ils cachent cependant quelques ambiguïtés. Nous en relevons deux :

i) Qu'entend-on par la "demande"? Au vue de la section 3.1, les sources d'appel sont fort nombreuses et plusieurs d'entre elles ne nécessitent pas une intervention immédiate. Doit-on appliquer, par exemple, la règle du 95% de couverture en 10 minutes à toutes les demandes, les plus urgentes comme les autres? De même, doit-on les appliquer dans toutes les sous-régions et à tout moment du jour ou suffit-il qu'elles soient respectées en moyenne?

ii) Qu'entend-on par une "couverture" en x minutes? La couverture est souvent évaluée en termes de "temps de réponse". C'est là un critère beaucoup plus facilement mesurable que celui du nombre de vies humaines sauvées par exemple. On s'entend généralement pour dire que plus le temps de réponse est court meilleure est la qualité du service et plus on pourra sauver des vies lorsqu'elles sont en danger. Mais même cette expression "temps de réponse" recèle ses ambiguïtés puisque plusieurs définitions sont possibles.

Dans la prochaine section, nous reproduisons le déroulement chronologique d'une intervention, depuis l'incident jusqu'au retour en disponibilité de la ressource affectée (supposée la seule ici). Nous insistons sur les définitions possibles du temps de réponse.

4.3 Chronologie d'une intervention

L'efficacité d'un système d'urgences médicales devrait se refléter à toutes les étapes menant à la réponse à un appel. À chacune de ces étapes, l'administration locale doit faire des choix judicieux pour améliorer l'efficacité globale du système.

Voici donc, pour un système général, la suite chronologique de ces étapes avec les choix qui s'offrent aux autorités, les définitions possibles du temps de réponse et les choix faits par Urgences santé.

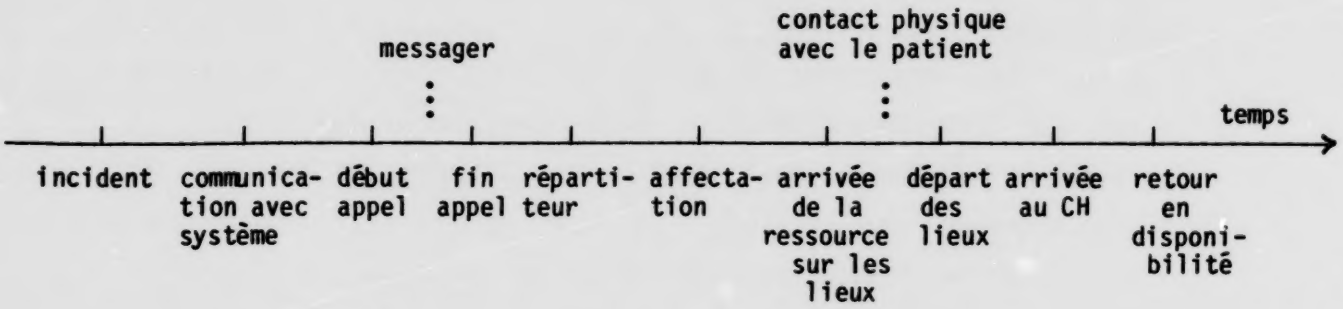

Figure 1 : Chronologie d'une intervention

incident : Ce moment n'existe pas toujours en fait. La décision de recourir au service ambulancier peut être instantanée (suite à une blessure, un accident de la route), mais elle peut également être l'aboutissement d'un long processus au cours duquel l'état du patient s'est progressivement dégradé ou ne s'est pas amélioré. Dans tous les cas, cet instant (s'il existe) nous est inconnu. Bien que selon nous c'est à partir de ce moment que devrait être calculé le temps de réponse, en pratique c'est impossible.

communication avec le système ambulancier : Nous la distinguons du début de l'appel pour tenir compte du fait que toutes les lignes peuvent être occupée et que le demandeur peut devoir attendre. À defaut de connaître le moment de l'incident c'est à partir de ce moment que le temps de réponse devrait être calculé pour refléter toute attente. À ce premier niveau de décision, le système d'urgence doit déterminer le nombre de préposés à la réception des appels.

début de l'appel : C'est l'instant où le premier intervenant amorce la communication avec le patient et prend le patient en charge. Le deuxième niveau de décision est le suivant :

- le personnel répondant n'a aucune qualification médicale. Il s'empresse de demander l'adresse de l'interlocuteur et cette information est acheminée au répartiteur qui y dépêche une ressource;
- le personnel répondant est médicalement qualifié (c'est le cas chez Urgences santé). Il demande l'adresse mais doit également recueillir des informations de nature médicale pour associer une priorité à l'appel et déterminer la(es) ressource(s) à affecter. L'information est alors communiquée au répartiteur (via un messager possiblement) ce qui n'empêche pas le répondant de poursuivre l'appel et de prodiguer des conseils à l'interlocuteur.

Dans un cas comme dans l'autre, on peut utiliser des outils informatiques pour obtenir automatiquement l'adresse de l'interlocuteur, par exemple via les banques de données de la compagnie de téléphone. C'est ce que fait depuis peu Urgences santé grâce au service 9-1-1 de Bell Canada. Ces banques de données associent l'adresse du payeur à un numéro de téléphone ce qui s'avère être l'adresse où est situé ce téléphone dans environ 95% des cas.

messager : Il peut être humain comme chez Urgences santé mais le tout peut être informatisé. Il assure le lien entre le répondant à l'appel et le répartiteur en transmettant la fiche comportant les données de l'appel.

répartiteur : Le répartiteur n'affecte pas nécessairement une ressource au moment où il reçoit la fiche-appel, et ce pour diverses raisons :

i) le répartiteur peut avoir à traiter d'autres appels au moins aussi urgents avant de traiter celui-ci. S'il n'y a pas de priorité associée aux appels alors ces appels sont simplement des appels étant survenus plus tôt. Sinon ce sont les appels au moins aussi prioritaires survenus avant, ou ceux plus prioritaires qui surviendront après et alors qu'aucune ressource n'a encore été affectée pour le présent appel.

ii) le répartiteur ne dispose pas de ressource pour répondre à l'appel. Cela peut vouloir dire qu'effectivement toutes les ressources sont occupées à répondre à d'autres appels. Mais cela peut aussi vouloir dire que l'appel n'étant pas de première urgence et le nombre de ressources libres étant faible, on juge qu'il est préférable de garder ces ressources pour d'éventuels appels plus urgents qui surviendraient à court terme. On parle alors de stratégie avec un nombre plancher fixé a priori.

affectation : C'est à ce moment qu'une ressource est affectée et qu'Urgences santé commence à calculer le temps de réponse. On omet ainsi, selon nous, tout un laps de temps où des décisions qui influencent le délai total et ont donc un impact direct sur la qualité du service reçu par le patient sont prises. Il est à noter, de plus,

affectation (suite) : qu'affectation ne signifie pas nécessairement démarrage immédiat du véhicule. L'ambulance peut en effet être immobilisée sur un terrain privé d'où elle doit possiblement sortir dans un fort trafic et en faisant un demi-tour.

arrivée sur les lieux : Ce moment est celui où le véhicule affecté s'immobilise à l'adresse de l'appel. S'il s'agit d'un accident de circulation les ambulanciers sont immédiatement en contact avec le demandeur, mais si ce dernier habite un édifice à appartements, il y a un délai pour le localiser et le contact avec le patient n'est pas instantané. C'est ce contact qui selon nous devrait compléter le calcul du temps de réponse à l'appel. Mais encore faut-il que le personnel ait les qualifications pour vraiment intervenir. Dans le cas contraire, ce n'est qu'au CH où le patient est véritablement pris en mains et ce n'est qu'à ce moment que la spécification du temps de réponse devrait être complétée.

départ des lieux : Sur les lieux, les ambulanciers peuvent, selon leur formation poser des actes médicaux. Dans certains cas, cela est suffisant et aucun transport vers un CH n'est requis. Dans d'autres cas, un transport vers un CH est nécessaire. Il est également possible que devant la situation, l'ambulancier décide que des ressources supplémentaires soient nécessaires, par exemple un médecin ou d'autres ambulances. Après avoir examiné le patient, les ambulanciers réévaluent son état. Selon l'urgence du cas, le trajet vers le CH s'effectuera à haute vitesse avec gyrophares et sirènes ou à vitesse normale.

arrivée au CH : Notons tout d'abord que ce CH n'est pas nécessairement celui qui est le plus près de la localisation du patient. La préférence du patient, des ambulanciers, la nature du malaise et l'achalandage des urgences sont autant de facteurs qui influencent ce choix. Rendu au CH, le patient est pris en charge par le personnel sur place, mais les ambulanciers peuvent avoir à transporter le patient dans le CH, remplir des formulaires, etc.

retour en disponibilité : En principe, la ressource est disponible dès sa sortie du CH, l'appel précédent complété. Cela peut cependant varier selon l'infrastructure du système d'urgence. Ainsi, chez Urgences santé, il est vrai que les ambulanciers se rapportent au répartiteur dès leur sortie du CH et ce dernier décide alors du poste où ils devront se diriger. Tout au long de ce trajet, ils sont considérés comme disponibles pour répondre à un appel. Pour d'autres systèmes, chaque ambulance est affectée à une caserne et à la fin d'un appel elle retourne toujours à cette même caserne et ce n'est qu'au moment où il y arrive qu'elle est considérée de nouveau disponible.

Plusieurs moments clés caractérisent un appel. Cela offre autant de possibilités différentes pour définir le temps de réponse. Selon la définition retenue, cependant, il sera plus ou moins facile d'atteindre les objectifs que l'on s'est fixés. C'est pourquoi il faut être prudent lorsque l'on analyse les performances d'un système d'urgence donné et bien connaître la signification des statistiques utilisées.

Dans le chapitre qui suit nous scrutons chacun des processus qui composent la problématique d'un système ambulancier.

CHAPITRE 5
LA PROBLÉMATIQUE

Maintenant que nous connaissons mieux quels sont les éléments en jeu, nous allons tenter de définir un certain nombre de problèmes que rencontrent les responsables d'un système d'urgence. Nous allons également mettre en évidence comment ces problèmes sont inter-reliés et s'influencent mutuellement.

Mentionnons d'abord brièvement les problèmes que nous avons choisis de ne pas traiter. Ils sont surtout d'ordre médical et nous les avons exclus de notre projet. À titre d'exemple, citons:

- le matériel qui devrait équiper les ambulances et les véhicules des médecins;
- la formation des intervenants, i.e. les répondants aux appels, les ambulanciers, etc.;
- les objectifs de qualité de service choisis. Nous nous sommes bornés à soulever les ambiguïtés inhérentes aux termes utilisés pour permettre aux autorités locales de faire un choix mieux éclairé à cet égard.

Nous présentons maintenant le contexte dans lequel nous avons choisi de travailler. Nous ne nous sommes pas attaqué à tous les éléments qui le constituent. Certains n'ont pas du tout été touchés, d'autres effleurés, pour d'autres enfin nous avons apporté une réponse assez complète. Schématiquement, ce contexte peut être décrit par la figure 2.

Nous reprenons ci-dessous chaque item apparaissant à la figure 2 et exposons en quoi il consiste, c'est-à-dire quels problèmes il comporte et quelles sont les solutions possibles.

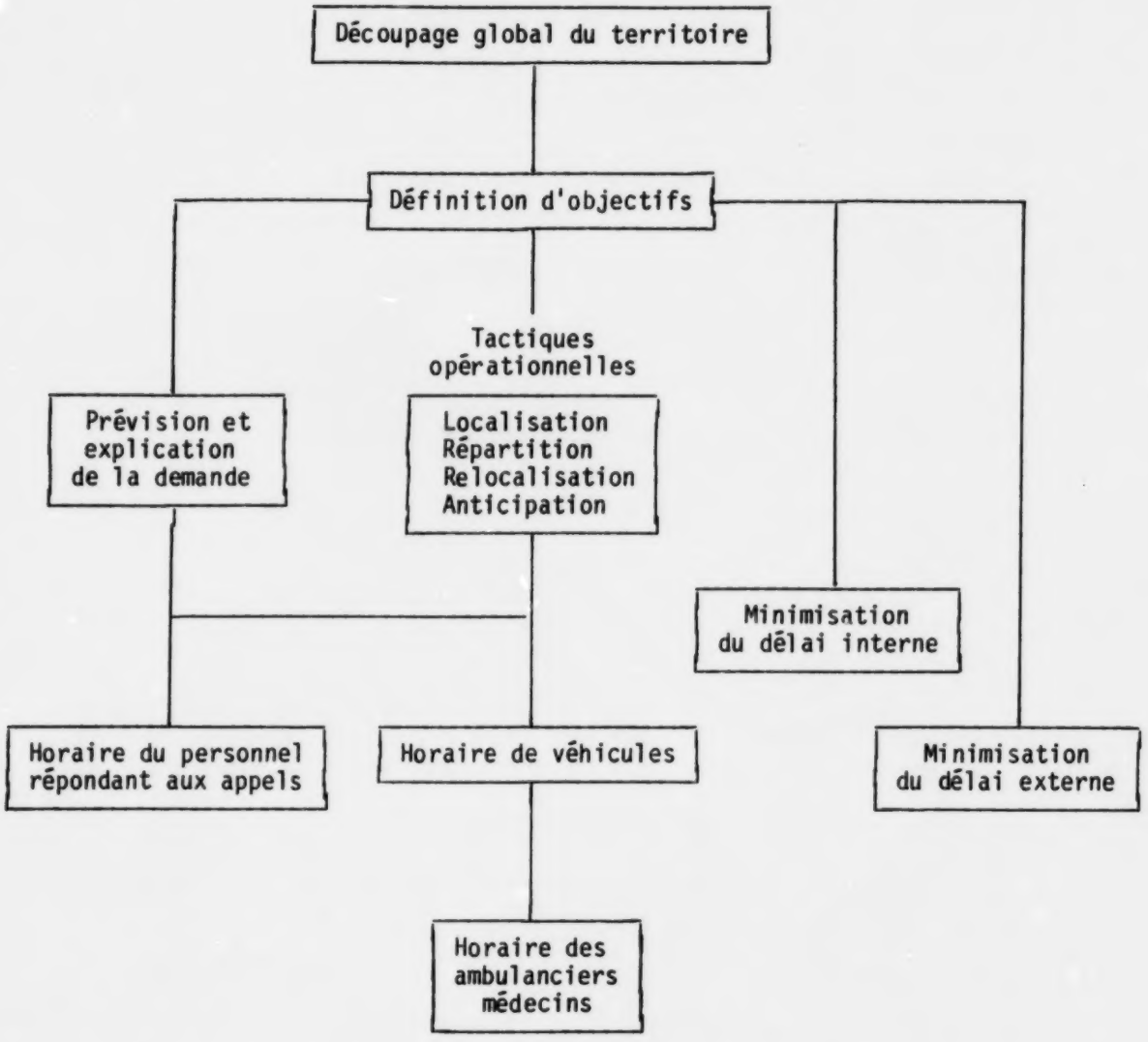

Figure 2 : Problématique d'un système ambulancier

5.1 Découpage global du territoire

Le problème peut s'énoncer comme suit : étant donné une région, comment devrait-elle être divisée en unités administratives disjointes de telle sorte que chaque unité fonctionne indépendamment des autres. On peut également considérer le nombre total de ressources disponibles et en exiger une répartition parmi les unités territoriales.

Dans notre contexte, cela pourrait vouloir dire : étant donné la région de Montréal (i.e. la rive Nord, l'île Jésus, l'île de Montréal et la rive Sud) quel est le meilleur découpage possible de l'ensemble de cette région en unités indépendantes? Bien sûr, la réponse à cette question dépend des objectifs que l'on s'est fixés au niveau de chaque région.

5.2 Définition d'objectifs

C'est là un élément clé. Tout le reste du développement vise précisément à atteindre les objectifs fixés. Ils sont avant tout spécifiés à partir de considérations médicales mais également par des considérations financières et législatives. Ils reflètent les préoccupations des autorités et sont le plus souvent multiples et parfois "antagonistes".

Tel que mentionné précédemment (voir section 4.1) nous n'avons pas fait d'étude systématique sur les objectifs qu'il est possible voire souhaitable de définir. Les motivations étant avant tout d'ordre médical, elles débordent notre domaine de compétence. Nous avons simplement fait un tour d'horizon des objectifs couramment utilisés, des exigences des lois, tout en faisant ressortir les ambiguïtés inhérentes à certains concepts, notamment à ceux de "couverture" et de "temps de réponse".

Une fois les objectifs définis, on peut considérer que la question majeure qui se pose est la suivante :

"Combien de véhicules devraient être mis en service à chaque moment du jour et quelle gestion de ces véhicules devrait être faite pour atteindre nos objectifs?"

Cette question comporte évidemment plusieurs facettes. Ainsi la gestion des véhicules a trait à leur localisation, la politique de répartition et à leur relocalisation (quand, comment?). Le nombre de véhicules doit être spécifié par catégorie (s'il y a lieu). Il faut également envisager une variation du nombre de véhicules en service, synchronisée avec celle de la demande.

Tous les blocs de la figure 2 qui apparaissent sous la définition d'objectifs aident à répondre plus ou moins directement à cette question.

5.3 Prévision et explication de la demande

Il va de soi que le nombre de véhicules à mettre en service dépend directement du niveau d'appel. Plus le taux d'appel est élevé plus les effectifs devront être nombreux pour permettre d'atteindre les mêmes objectifs mais la relation entre les deux n'est pas linéaire. Or les appels constituent un phénomène aléatoire à la fois spatial et temporel. Qui plus est, le taux, disons horaire, d'appel n'est pas connu à l'avance et varie en général au cours d'une même journée, de jour en jour, de saison en saison, etc. Il est donc primordial de se munir d'un outil de prévision de la demande. Il est de plus souhaitable de bénéficier d'un modèle d'explication de la demande pour identifier les causes des variations du niveau d'appels. Ces causes peuvent être de nature socio-économique, climatique, etc.

Chez Urgences santé, l'horaire est le même tout au long d'une saison. On pourrait se contenter de se prémunir contre les heures de pointe et mettre en service le même nombre de véhicules tout au long du jour. À Montréal, tout au moins, cela entraînerait un gaspillage énorme de ressources puisque le taux d'appel varie du simple au double entre les périodes les

moins et plus occupées de la journée. Urgences santé a plutôt choisi de tenter d'épouser au mieux la courbe de la demande en considérant un découpage horaire.

5.4 Horaire du personnel répondant aux appels

Le traitement d'un appel ne débute que lorsqu'un répondant prend cet appel. Il peut arriver que tous les répondants soient occupés et que des appels s'accumulent en file d'attente. Cette situation n'est pas souhaitable puisqu'elle retarde d'autant l'intervention de ressources sur les lieux de l'appel. Or le temps de réponse est souvent des plus critique. Il importe donc que le personnel répondant soit en nombre suffisant pour limiter au minimum ces délais.

Nous avons là, en beaucoup plus simple, un système semblable à celui des véhicules d'intervention. Il faut déterminer le nombre de répondants qui devraient être mis en service à chaque moment du jour pour satisfaire aux objectifs définis.

La demande ici est un peu différente de celle à laquelle font face les véhicules d'urgence puisqu'elle inclut, en plus, tous les appels qui ne génèrent aucun transport et qui sont réglés directement par le répondant (voir section 3.1). D'autre part, la gestion des ressources (les répondants ici) est beaucoup plus simple qu'avec les véhicules d'urgence. Il n'est pas question de localisation, ni de relocalisation. Nous n'avons pas à nous préoccuper de la dimension spatiale de la demande. Seuls la durée de l'intervention de ces répondants, le taux d'appel et le nombre de répondants a un effet sur la présence éventuelle de file d'attente. C'est donc un système plus simple mais le seul fait que le taux d'appel varie complique déjà la question.

5.5 Horaire des véhicules

Une fois la demande identifiée par le modèle de prévision de la demande, elle est convertie en nombre de véhicules à mettre en service.

Cette conversion constitue un processus complexe qui tient compte à la fois des objectifs à atteindre, de la stratégie opérationnelle adoptée, de la distribution du temps de service des véhicules d'urgence, etc. Pour l'instant, supposons que cette conversion a été complétée. On obtient un graphique semblable à celui de la figure 3. Pour simplifier, supposons que la couverture requise est la même pour les sept jours de la semaine de toutes les semaines de la saison considérée.

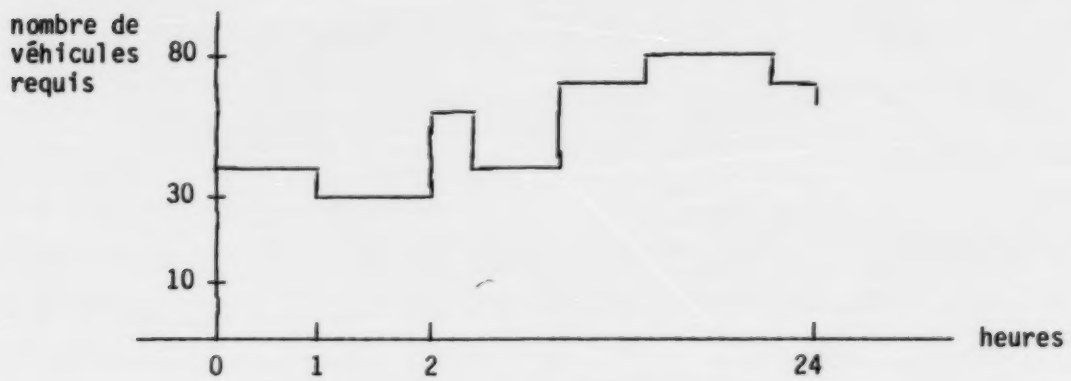

Figure 3 : Demande horaire de véhicules

Il faut maintenant confectionner un horaire de véhicules qui, globalement, engendre la couverture désirée. Le tout doit être réalisé en respectant un ensemble de contraintes, issues pour la plupart de la convention collective des ambulanciers. En voici un échantillon tiré du cas d'Urgences santé :

- nombre maximum de véhicules disponibles;
- durée permise des pièces de travail (une pièce de travail, par exemple de 8h00 à 16h00, deviendra éventuellement la journée de travail d'une équipe d'ambulanciers);
- écart minimum entre deux utilisations successives d'un même véhicule (pour en faire l'entretien);
- intervalle de temps à l'intérieur duquel les pièces de travail peuvent débuter (jamais la nuit);

- répartition des heures de service au pro rata des effectifs entre les diverses compagnies possédant des ambulances;
- nombre de véhicules pouvant débuter simultanément une pièce de travail (cette contrainte tient au nombre limité de fréquences disponibles pour se rapporter aux répartiteurs lorsqu'un quart débute);
- délai minimum entre deux débuts de quart successifs.

5.6 Horaire des ambulanciers, médecins

Une fois l'horaire des véhicules complété, il reste à préciser celui du personnel qui opère ces véhicules. Déjà à l'étape précédente nous avons tenu compte de certaines des contraintes que doit respecter l'horaire du personnel (la durée des quarts, par exemple). D'autres contraintes s'ajoutent, comme par exemple

- nombre total d'heures de travail sur deux semaines;
- favoriser une certaine disposition des journées de congé (consécutives, la fin de semaine de préférence);
- favoriser des quarts semblables (c'est-à-dire débutant aux mêmes heures) pour une même personne;
- écart minimal entre deux quarts successifs d'une même personne.

Dans les sections 5.7 et 5.8 qui suivent nous introduisons les notions de délai interne et externe. Tout délai a avant tout une influence directe sur la qualité des soins offerts et le premier à en souffrir c'est le patient. Mais même lorsque le patient est stabilisé et hors de danger, le temps perdu prive le système d'une ressource qui pourrait servir à traiter d'autres patients.

5.7 Minimisation du délai interne

Par délai interne nous entendons ceux se déroulant entre le début de l'appel et l'affectation d'un véhicule. On le qualifie d'interne puisque toutes les opérations au cours de ce laps de temps se déroulent au poste de contrôle du système ambulancier.

Par souci d'efficacité on désire limiter au minimum ce délai. Voyons ses composantes :

i) **attente téléphonique :**

Il s'agit du temps que met le demandeur à obtenir la communication avec un des répondeurs d'Urgences santé. Il y a délai lorsque tous les répondeurs sont occupés. Le remède c'est d'ajuster le mieux possible le nombre de répondeurs avec la demande, tout en respectant des contraintes de coût et de convention collective.

ii) **début appel-messager :**

Il correspond au temps nécessaire pour recueillir les renseignements minimaux avant de pouvoir les transmettre au messager. Ces renseignements sont

- l'adresse d'où émane l'appel. Pour accélérer ce processus, Urgences santé utilise maintenant la banque de données de Bell Canada qui associe une adresse à un numéro de téléphone. Deux problèmes demeurent :
  - la fiabilité de la banque et sa mise à jour;
  - la ressource n'est pas toujours requise sur les lieux d'où émane l'appel.
- l'établissement d'un diagnostic et des ressources à envoyer.

Chez certains organismes cette dernière étape n'a pas lieu (ex. : Toronto) on dépêche systématiquement une ressource pour chaque appel. On limite cette partie du délai interne mais au risque de congestionner le système et d'augmenter le temps qui s'écoule entre le moment où le répartiteur prend connaissance de l'appel et l'arrivée sur les lieux d'une ressource.

L'établissement du diagnostic peut être accéléré par un système-expert qui guide rapidement l'infirmier(e) vers les conclusions à tirer. De même, si un système informatique le renseignait sur les ressources disponibles et leur localisation, il pourrait mieux évaluer les ressources qu'il est raisonnable d'envoyer. A la limite, cette décision

pourrait faire partie intégrante du système-expert. Ainsi, si en théorie il faut un médecin mais que la ressource la plus près de l'appel est trop éloignée, on préfèrera envoyer une ambulance plus proche qui amènera le patient rapidement à un CH.

iii) **messager :**

Le messager achemine l'information relative à un appel du répondeur jusqu'au répartiteur. Chez Urgences santé c'est un humain qui remplit cette fonction. Il est certain que le transfert se ferait plus rapidement si tout était informatisé, sans intermédiaire.

iv) **l'affectation :**

Le répartiteur a généralement devant lui plusieurs fiches correspondant à autant d'appels non encore affectés. Si le système utilise des priorités, il les gère selon la politique établie. Divers facteurs peuvent influencer la durée du traitement d'une fiche et donc le délai interne :

- autres tâches

  Le répartiteur peut avoir à se consacrer à d'autres tâches. Ainsi, chez Urgences santé les répartiteurs ont également la responsabilité, de contrôler,

  • les débuts et fins de quart
  • les débuts et fins de dîner
  • les relocalisations.

  Il doivent aussi guider une ressource qui aurait de la difficulté à trouver une adresse. De plus, ils n'interrompent pas une tâche en cours pour débuter une affectation.

- localisation

  C'est tout le travail de recherche nécessaire à la réalisation d'une affectation :

  • localisation de l'adresse : actuellement le répartiteur doit, le cas échéant, chercher manuellement dans un répertoire des cartes de la région de Montréal. Ce serait beaucoup plus rapide s'il bénéficiait d'un outil graphique réalisant cette recherche dès que l'adresse est connue (c'est-à-dire avant même l'intervention du répartiteur) et pourrait l'afficher au besoin sur écran.

- localisation des ressources les plus près : c'est-à-dire identifier parmi les types de ressource requis les véhicules les plus près de l'appel. Encore là, comme un véhicule libre peut être en mouvement il peut être difficile pour le répartiteur d'identifier le mieux placé. Un outil graphique associé à une localisation automatique des véhicules sur le territoire accélèrerait grandement cette recherche. A cet effet, Urgences santé a mis sur pied un projet CAD (Computer Aided Dispatch) pour réaliser ce processus.

- affectation

Après avoir identifié les ressources pour répondre à un appel, il doit les avertir de leur tâche. Pour ce, il doit communiquer à tour de rôle avec elles et leur énumérer l'information contenue sur la fiche appel, à savoir

- l'adresse
- le diagnostic
- s'il s'agit d'un déplacement à petite ou grande vitesse.

Tout cela prend un certain temps qui pourrait être mieux utilisé. Il faudrait alors que cette information soit automatiquement transmise de l'ordinateur central aux ressources qui bénéficieraient alors d'un mini terminal leur permettant d'afficher ces lignes d'information.

5.8 Minimisation du délai externe

Par délai externe nous entendons la somme des délais ayant court à partir de l'affectation et qui sont indépendants de la tactique opérationnelle générale. Ce délai est qualifié d'externe car il se produit à l'extérieur du centre de contrôle des opérations. Tout comme pour le délai interne on désire le réduire au mininum. Dans l'ordre chronologique il est constitué des items suivants :

i) **ambulancier hors du véhicule :** dans certains systèmes (ex. Toronto) les ambulanciers sont à la caserne hors de leur véhicule lorsqu'ils sont affectés à un appel. Il faut un certain temps avant qu'ils soient prêts à démarrer.

ii) **la recherche de l'adresse avant même de démarrer :** chez Urgences santé les ambulanciers vont d'abord identifier le lieu de l'adresse avant de démarrer. C'est selon eux plus facile d'effectuer cette

recherche alors qu'on est immobilisé et d'éviter d'aller dans une mauvaise direction et de perdre encore plus de temps. Par un système de localisation informatisé, on pourrait faire automatiquement cette recherche et en communiquer le résultat (soit l'afficher sur un écran dont serait munie l'ambulance, soit lui communiquer la référence dans un atlas).

iii) **choix du parcours :** il existe en général plusieurs chemins pour aller d'un point A à un point B sur un réseau. L'ambulancier est intéressé à utiliser le plus rapide. Ce chemin idéal peut d'ailleurs varier selon le moment de la journée. De par son expérience, l'ambulance connaît les artères les plus congestionnées. S'il en est averti, il peut également tenir compte de certaines situations d'exception (un incendie ou un accident de la circulation qui bloque une rue). Cela ne veut toutefois pas dire qu'il emprunte toujours le meilleur chemin.

iv) **démarrage :** chez Urgences santé les postes d'attente ne sont pas sur le réseau routier même. Les véhicules sont le plus souvent dans un stationnement (de centre d'achat, magasins, ...). Il faut un certain temps pour en sortir, même avec gyrophares et sirènes. Si ce lieu est par exemple sur un grand boulevard avec terre-plein et que l'on veut accéder aux voies de l'autre côté du terre-plein, il faut faire un virage en U. En général, la solution à ce problème réside dans le choix des postes d'attente qui devraient permettre d'accéder à toutes les directions facilement.

v) **bris mécanique :** dans un tel cas, une autre ressource doit prendre la relève.

vi) **contact avec le patient :** une fois arrivé à l'adresse prévue l'ambulancier immobilise son véhicule et se met à la recherche du patient. Lorsque le patient se trouve à l'extérieur (ex. : accident de la route) cette recherche n'est pas bien longue. Dans le cas contraire, on doit chercher la localisation exacte (l'appartement, le local, etc.). Même si l'ambulancier ne perd pas de temps cela peut être long en raison de moyens d'accès difficiles : longue marche à faire, escalier à gravir, etc.).

vii) **CH de destination :** supposons maintenant qu'un transport vers un CH soit nécessaire. Le choix du CH dépend à la fois des goût et antécédants du patient, de la préférence des ambulanciers, de la proximité des CH, du malaise à traiter (certains CH ne pouvant traiter certains cas), et de l'engorgement des urgences des CH. Il convient surtout ici d'adopter une politique claire et surtout de travailler en concertation avec tous les intervenants pour éviter à l'ambulancier des détours inutiles (par exemple d'aller à un CH et s'y voir refuser l'accès ou aller à un CH beaucoup plus loin que nécessaire). Le problème d'engorgement des urgences est particulièrement aigu à Montréal et il importe donc :

- de communiquer en tout temps avec les urgences pour connaître leur niveau d'occupation et éviter de s'y rendre pour rien;
- tenter de mieux répartir les patients entre les diverses urgences pour éviter d'emplir trop vite certaines qui, devenant non disponibles, peuvent occasionner par la suite un délai fatal.

En résumé, la coordination entre urgences, répartiteur et ambulanciers est primordiale.

viii) **temps passé au CH :** le travail qu'effectue l'ambulancier au CH est variable. Si les renseignements qu'il a à fournir sont partout à peu près les mêmes, la manipulation du patient, elle, varie. Il est plus long de transporter un patient dans un département que de le laisser à

l'urgence. Évidemment, plus ce temps est long, moins vite les ambulanciers redeviendront disponibles pour répondre à un autre appel. Il y a donc à cet égard un travail de sensibilisation à faire auprès des autorités des divers CH pour réduire au minimum ce temps passé au CH.

5.9 Tactiques opérationnelles

5.9.1 Introduction

Les délais internes et externes sont des sources potentielles de perte de temps dont la réduction dépend souvent d'outils technologiques permettant de diminuer le délai dans la transmission d'une information. La tactique opérationnelle représente tout l'autre champ d'intervention qui permet une meilleure utilisation des ressources. Dans cette section nous présentons l'éventail des stratégies possibles.

Par tactique opérationnelle, on entend l'ensemble des décisions prises relativement aux quatre questions suivantes :

i) Localisation : où localiser les postes d'attente?

Un poste d'attente est un lieu où on peut localiser une ambulance libre. Dans certains systèmes ce lieu peut être une caserne pouvant accomoder plusieurs véhicules et parfois être partagée avec le service des incendies. Dans d'autres contextes d'exploitation ces lieux sont désignés par une adresse, l'intersection de deux rues, le stationnement d'un centre d'achat, d'un restaurant, d'un hôpital. Les casernes sont des installations permanentes et on ne remet généralement pas en question leur localisation. Lors de la construction de nouvelles casernes cependant la question de localisation devient pertinente. Lorsque les postes d'attente ne sont pas des casernes on a beaucoup plus de flexibilité pour remettre en question les localisations retenues lorsque certains paramètres du système changent. On peut également vouloir varier les localisations au cours d'une journée en

fonction de la variation dans la demande, dans la congestion du réseau routier, etc. Lorsqu'on utilise une tactique de patrouille il n'y a pas de postes d'attente à proprement parler. Dans ce contexte, le problème de localisation revient à définir les régions patrouillées.

ii) Affectation : quelle(s) ressource(s) affecter à un appel?

Lorsque survient un appel, sa nature détermine les types de ressources requises. Parmi celles disponibles lesquelles envoyer? La réponse qui vient immédiatement à l'esprit est d'envoyer la ressource la plus près de l'appel pour chacun des types requis.

Dans la pratique les choses ne sont pas tout à fait aussi simples :

- dans certains systèmes le territoire est partitionné et on affecte systématiquement la ressource responsable du territoire d'où vient l'appel même si elle n'est pas la plus près;
- lorsque des ressources disponibles sont en mouvement il peut être à la limite plus long de s'enquérir de leur position pour déterminer la plus près de l'appel que d'affecter celle qui apparaît la plus près;
- théoriquement la règle "envoyer la ressource la plus près de l'appel" n'est généralement pas optimale [Carter, 1972], bien que très proche (environ 2% de l'optimum [Jarvis, 1981]). La majorité des systèmes utilise cependant cette règle simple et pratique.

iii) Relocalisation : quand et comment effectuer une relocalisation?

Pour les systèmes où aucune relocalisation n'est permise, la question est vite réglée. Pour les autres il faut spécifier :

- quelle(s) condition(s) nécessitent une relocalisation des véhicules disponibles. Intuitivement, plus on relocalise souvent mieux c'est.

Mais chaque relocalisation demande beaucoup d'effort car il faut communiquer avec chacun des véhicules relocalisés encombrant ainsi les ondes et les répartiteurs. L'utilisation d'un système informatisé permet toutefois de réduire grandement ces problèmes. Le nombre de déplacements possibles d'une ambulance à l'intérieur d'une période de temps donné reste limité. De plus, le gain réel est parfois faible et n'en vaut pas la peine;

- une fois la décision prise de relocaliser il faut identifier :
  - quels nouveaux postes d'attente devront être couverts
  - quels déplacements les véhicules présentement disponibles devront effectuer pour accomplir cette tâche tout en atteignant certains objectifs comme la minimisation du nombre, de la durée et de la distance associée à ces déplacements.

Ces questions sont fort complexes. Il faut en particulier tenir compte du fait que durant la relocalisation des appels peuvent survenir. D'une part les déplacements devront anticiper le mieux possible ces appels et d'autre part à la fin de la relocalisation on se retrouvera dans une configuration qui est inconnue à l'avance.

iv) Anticipation : comment tirer profit de notre connaissance de certains événements à venir?

Certains événements futurs sont connus de façon sûre : le début ou la fin d'un quart de travail ou d'un repas à une heure donnée. Les connaissant on peut tenter d'en tirer profit. Ainsi si un poste d'attente est occupé par une seule ressource et que cette dernière termine son quart de travail à brève échéance, il serait plus prudent de couvrir le poste par une deuxième ressource (si possible) en anticipation du départ de la première. Notons qu'en pratique on n'est jamais à l'abri d'un retard et que les événements futurs connus de façon sûre sont plutôt rares.

D'autres événements futurs sont connus mais avec une certaine probabilité seulement. Doit-on en tenir compte et si oui comment? Par exemple, une ressource R est occupée à répondre à un appel depuis cinquante minutes et les données du passé indiquent alors une probabilité conditionnelle de 0,95 pour un retour en disponibilité dans les trois prochaines minutes. Connaissant le CH où cette ressource redeviendra disponible devrait-on en tirer profit, par exemple on ne couvrant pas le poste d'attente libre le plus près de ce CH sachant que R pourra le faire très bientôt? Mais avant de mettre en application une telle politique, il faut être très prudent. En tentant une telle implantation, Urgences santé s'était rendu compte que le temps passé au CH par les techniciens ambulanciers était très variable et influencé par plusieurs facteurs incontrôlables.

5.9.2 Énumération des stratégies de localisation/relocalisation

Nous avons arbitrairement réparti les stratégies possibles en deux classes et subdivisé ces classes comme suit :

- celles où l'on ne permet pas de relocaliser les véhicules
  - districts indépendants
  - districts dépendants
- celles où l'on permet les relocalisations
  - postes d'attente fixés
  - postes d'attente variables
  - patrouille

A) Sans relocalisation

Dans ce cas le scénario est le suivant : des casernes ayant chacune ses effectifs sont aménagées sur le territoire. Lorsqu'une ambulance est affectée à un appel, elle y répond puis revient toujours à la même caserne à laquelle elle est rattachée.

i) **districts indépendants**

Dans ce cas, les casernes sont indépendantes en ce sens que le territoire est partitionné en districts et chaque caserne a la responsabilité exclusive d'un district dans le sens où il n'y a aucune coopération entre les casernes. S'il arrivait que les ambulances d'un district soient toutes occupées et qu'un autre appel survienne dans ce même district alors on devrait attendre que l'une des ambulances occupées du district se libère. Il n'est pas question de demander aux casernes avoisinantes de nous dépanner. Il n'y a aucun choix possible au niveau des règles d'affectation : c'est toujours une ambulance de la caserne du district qui répond aux appels du district.

Le problème dans ce contexte se résume donc à :
- où bâtir les casernes?
- comment définir les districts?
- combien mettre d'ambulances dans chacune des casernes?

Ce type de tactique est utilisé à Toronto.

ii) **districts en coopération**

Cette fois-ci les casernes peuvent s'entraider. Chaque ambulance reste rattachée à une seule caserne et y retourne à la fin de chaque appel auquel elle répond. Mais si une caserne se vide alors son territoire est couvert par les casernes avoisinantes. En fait il n'est plus vraiment nécessaire de définir de districts comme tel. On pourrait, par exemple, adopter une règle de répartition ("dispatch") qui envoie toujours la ressource disponible la plus près de l'appel.

Dans ce contexte il faut donc décider :
- où localiser les casernes?
- combien mettre de ressources dans chacune?
- quelle règle de "dispatch" mettre en vigueur?

Ce type de tactique est utilisé à Boston et Austin.

B) Avec relocalisation

i) **postes d'attente fixes**

On peut imaginer ici que l'on a encore des casernes, c'est-à-dire des entités immuables mais cette fois-ci une ambulance n'est plus nécessairement rattachée à une seule caserne. On dispose plutôt d'une flotte de véhicules où chacun peut occuper, selon les besoins, l'un ou l'autre des postes d'attente à divers moments de son quart de travail. Le but de la relocalisation est d'assurer en tout temps un déploiement satisfaisant des ressources disponibles. À titre d'illustration, supposons qu'on ait douze postes d'attente distribués schématiquement comme dans la figure 4, et douze ambulances. Supposons, de plus, que coup sur coup six appels surviennent qui nous privent des ressources 1 à 6. Il ne reste donc plus que six ressources, toutes concentrées à l'est du territoire. Si le prochain appel provient du district 2, par exemple, alors le temps de réponse de la ressource la plus près (l'ambulance 8) risque d'être plutôt long. Pour se prémunir contre cette éventualité on redéploie (et on aurait pu le faire avant que les six ressources disparaissent) les ressources restantes de manière à mieux répondre au prochain appel, d'où qu'il provienne.

•1	•4	•7	•10
•2	•5	•8	•11
•3	•6	•9	•12

Figure 4 : Relocalisation avec postes d'attente fixes

ii) **postes d'attente non permanents**

C'est essentiellement la même stratégie qu'en i) ci-dessus. Cette fois-ci on n'a pas vraiment de casernes mais plutôt des lieux (des

intersections, un terrain de stationnement, etc.) qui peuvent changer selon le moment du jour, les conditions météo, la saison, etc. C'est le type de stratégie utilisé à Montréal.

iii) **la patrouille**

Si l'on pousse le principe de la relocalisation à la limite on obtient des véhicules qui n'occupent plus des postes d'attente fixés mais plutôt qui patrouillent des territoires variables selon le nombre d'ambulances disponibles. C'est vraiment le cadre le plus souple, très familier dans le contexte policier, mais nous ne connaissons aucune ville qui utilise ce type de tactique pour des ambulances.

En fait, à partir des quatre blocs décrits on peut imaginer de multiples variantes. Ainsi, Urgences santé utilise une variante de la stratégie avec relocalisation où les postes d'attente ne sont pas permanents. Vingt-cinq postes sont utilisés. Ils sont localisés de manière à ce que certains objectifs soient atteints lorsqu'ils sont tous occupés. Toute la stratégie vise donc à ce que chacun d'eux soit occupé en tout temps. Lorsqu'un poste est à découvert, on relocalise des véhicules en vue d'occuper ce poste. Lorsque plus de vingt-cinq ressources sont libres, certains postes sont occupés par plus d'un véhicule.

Les types de tactique ont été présentés dans l'ordre croissant du degré de flexibilité. Or, puisque l'ensemble de solutions réalisables grandit avec le degré de flexibilité, il s'en suit que vis-à-vis d'un même objectif, un système plus contraint ne pourra faire mieux qu'un système moins contraint. Ainsi, permettre de relocaliser sa flotte ne peut qu'améliorer les performances. En ce sens, Urgences santé jouit présentement d'une grande liberté. Elle n'est pas confinée à utiliser des casernes et elle peut effectuer des relocalisations. Nous sommes donc particulièrement

intéressés aux systèmes les moins contraints. Mais le prix à payer pour relocaliser est le nombre de manoeuvres supplémentaires qu'il faut effectuer, ce qui congestionne les ondes de communication entre répartiteurs et ambulanciers et pendant ce temps les répartiteurs ne font pas de dispatching. Tout cela en vaut-il la peine? Il y a donc intérêt à mesurer l'avantage relatif de chaque tactique.

Maintenant que la problématique est définie, nous présentons au chapitre 6 les éléments sur lesquels nous avons travaillé et les résultats que nous avons obtenus.

CHAPITRE 6

RECHERCHE BIBLIOGRAPHIQUE ET SOLUTIONS RETENUES

6.1 Introduction

Parmi tous les problèmes décrits au chapitre 5, il nous a fallu faire un choix. Nous n'avions ni le temps ni les ressources pour les attaquer tous. Dans ce chapitre nous revenons sur chacun de ces problèmes pour indiquer si nous nous y sommes attardés. Dans l'affirmative, nous faisons un bref survol de la littérature et nous décrivons et justifions l'approche retenue. Finalement, nous référons le lecteur intéressé à d'autres de nos publications pour plus de détails quant à la méthodologie, aux résultats et au travail qu'il reste à faire.

Nous avons concentré nos efforts dans trois directions principales : confection d'horaires, prévision et explication de la demande, génération et analyse de tactiques opérationnelles.

6.2 Découpage global du territoire

Nous n'avons rien fait à cet égard. Notons que le problème ne s'est pas (encore) posé à Montréal.

6.3 Définition d'objectifs

Tel que mentionné précédemment, nous n'avons pas fait d'étude exhaustive de ce thème puisqu'une bonne partie des décisions à cet égard est l'ordre médicale et déborde notre champ d'expertise. L'essentiel de notre réflexion est décrit aux sections 4.1 et 5.2.

Dans le cours du projet, nous avons fait part de nos réflexions à Urgences santé en insistant particulièrement sur la notion d'équité. Urgences santé a formé un comité chargé d'étudier la question et modifier au besoin les objectifs jusqu'alors définis. Cette étude a conduit à une redéfinition des objectifs qui s'expriment maintenant comme suit :

Minimiser le temps moyen de réponse (selon leur définition) en s'assurant que

Prob [t. rép. d'un appel STAT < 5 min.] > 0,95 et que

Prob [t. rép. d'un appel STAT < 10 min.] = 1,00

en conservant les mêmes objectifs qu'auparavant pour les appels de moindre priorité. Ces buts officieux étaient les suivants :

pour les appels de priorité STAT et 1 : répondre en au plus 5 min
pour les appels de priorité 2 et 3 : répondre en au plus 20 min
pour les appels de priorité 4 et 5 : répondre en au plus 60 min.

6.4 Prévision et explication de la demande

La prévision de la demande doit nous renseigner sur le niveau horaire de la demande à chaque saison. Pour ce faire, nous avons choisi d'utiliser un modèle de série chronologique ARIMA [Box, 1976]. Quant à l'explication de la demande, elle a pour but d'identifier des facteurs qui influencent le niveau journalier de la demande. Cette étude a peu d'impact direct sur la conception des horaires, mais elle peut aider à mesurer l'effet de divers facteurs tels une nouvelle tarification, les grèves du personnel, les jours fériés, les variations climatiques, etc.

Jusqu'ici seul l'aspect explication de la demande a été touché. Dans la suite de ce projet nous entendons aborder la prévision de la demande. Nous avons agi ainsi car nous disposons déjà au C.R.T. de l'outil approprié pour identifier les variables ayant la plus grande influence sur la demande. Il s'agit du modèle L-1.1 [Dagenais, 1980], [Liem, 1983]. Il a été récemment implanté, avec d'autres modèles, sur miniordinateur sous le nom de TRIO.

6.5 Horaire du personnel répondant aux appels

Nous n'avons rien fait à cet égard. Soulignons toutefois que la démarche à suivre pour y arriver est identique à celle conduisant aux horaires des ambulanciers et qu'il suffirait de modifications mineures aux logiciels existant pour bénéficier d'un tel outil.

6.6 Horaire des véhicules

Pour résoudre ce problème nous avons utilisé un logiciel développé au C.R.T., HASTUS-MACRO [Rousseau, 1978], [Blais, 1978], dans le cadre de la confection des horaires de chauffeurs d'autobus. Pour des raisons de taille d'espace mémoire, nous n'avons pu tenir compte de toutes les contraintes à la fois. Aussi, nous n'avons permis que des départs aux heures (en limitant ce nombre de telle sorte que la contrainte puisse être respectée). Par la suite, cette solution est traitée par un autre logiciel, DECAL, qui décale les quarts de travail de manière à respecter la contrainte violée (et tout en s'assurant de ne pas en violer d'autres, notamment l'écart minimum entre deux pièces d'un même véhicule).

Par le passé, l'horaire des véhicules chez Urgences santé était fait manuellement. Ils possèdent maintenant un logiciel directement inspiré de nos travaux. En plus d'accéler considérablement ce processus, il génère des solutions plus près de la couverture désirée.

6.7 Horaires des ambulanciers

La qualité de l'horaire du personnel dépend fortement de celui des véhicules. Par conséquent, les deux devraient en principe être faits simultanément. Nous avons choisi de les faire séquentiellement en raison de la complexité de chacun d'eux. On peut toutefois établir une structure de coût dans le premier algorithme qui facilite par la suite la construction du deuxième.

Un logiciel permettant la construction des horaires des techniciens ambulanciers est maintenant en place chez Urgences santé. Il distribue, dans un premier temps, l'ensemble des quarts de travail aux différentes compagnies ambulancières à l'aide d'une structure de coûts assurant un minimum de qualité aux horaires des ambulanciers.

Dans un deuxième temps, les quarts sont attribués à des équipes de travail en fonction de nombreux critères (débuter toujours à la même heure, ne pas travailler en trop grand nombre de jours consécutifs, etc.) à l'intérieur de chaque compagnie.

6.8 Minimisation des délais interne et externe

En ce qui concerne tous les éléments du délai interne, le projet initial prévoyait la conception d'un outil graphique d'aide à la répartition. Cet outil aurait grosso modo couvert l'item iv) de l'affectation (voir 5.7iv) en affichant le lieu de l'appel et les ressources disponibles les plus près en utilisant l'information fournie par le CAD.

Urgences santé ayant manifesté un grand intérêt pour le logiciel des horaires de véhicules et de personnel, nous avons poussé beaucoup plus loin que prévu leur développement, au détriment de cet outil graphique. Soulignons toutefois que d'autres projets du centre travaillent au développement d'un outil semblable qui pourrait être adapté au contexte d'Urgences santé.

6.9 Tactiques opérationnelles

C'est un des trois axes auxquels nous nous sommes consacrés et c'est certainement celui qui représente le plus grand défi. Nous avons d'abord voulu couvrir le plus large éventail possible de contextes, du plus contraint au moins contraint (voir section 5.9.2). Pour chacun d'eux nous visions à identifier la meilleure tactique possible. Pour générer des tactiques nous avons eu recours à des méthodes analytiques lorsque c'était possible mais également à des heuristiques.

Il nous fallait d'autre part un moyen d'évaluer les différentes tactiques. Ce besoin est d'autant plus critique que nous évoluons dans un contexte complexe où la santé du public est en jeu et où il ne saurait être question d'expérimenter en mettant des vies en danger. C'est pourquoi beaucoup d'énergie a été dépensée à la conception et à la réalisation d'un outil de simulation [Trudeau, 1987d], [Choquette, 1987]. Il nous permet de tester le comportement d'une tactique dans un environnement très proche de la réalité.

Avant d'aborder la revue de la littérature, nous énumérons les critères qui ont guidé notre choix à travers toutes les publications.

6.9.1 Critères de sélection d'une approche de solution

De manière générale, on peut définir une approche de solution par la donnée

- d'un modèle;
- d'un algorithme de résolution.

Cette approche est caractérisée par

- les hypothèses du modèle;
- les performances de l'algorithme (taille de la mémoire requise, temps d'exécution);
- la nature de l'approche, c'est-à-dire (dans notre contexte)
  - ou bien un outil décrivant les performances d'une stratégie fournie en input;
  - ou bien un outil recherchant une stratégie optimale à partir d'une stratégie initiale fournie en input.

L'outil idéal devrait poser des hypothèses les plus générales possibles (c'est-à-dire les moins contraignantes), être performant (rapide et pouvant traiter des problèmes de grande taille) et rechercher et identifier une stratégie optimale globale. Dans notre contexte cet outil n'existe pas et ne représentera pas l'aboutissement de nos recherches.

Passons chacune de ces caractéristiques en revue.

- **hypothèses du modèle** : lorsque l'approche est de nature analytique les hypothèses sont habituellement plus fortes puisque posées pour rendre un tel traitement du modèle possible, beaucoup plus que pour refléter la réalité. Il ne faut cependant pas s'arrêter à la seule vraisemblance des hypothèses. Il faut analyser la valeur réelle des solutions générées par de tels modèles. Les solutions peuvent malgré tout être intéressantes, surtout en l'absence d'autres outils.

- **performances de l'algorithme :** une de nos préoccupations est de pouvoir traiter des problèmes de la taille de ceux rencontrés à Montréal, c'est-à-dire comportant jusqu'à 80 ambulances simultanément en service. C'est là un problème de grande taille puisque ceux rencontrés dans la littérature ont presque tous moins de dix ambulances. De plus, les villes que nous connaissons utilisent rarement autant d'ambulances (ex. : 12 véhicules à Austin [Eaton, 1985], 9 véhicules à Boston [Brandau, 1983]). Nous sommes donc peu intéressés à des modèles sophistiqués qui ne peuvent résoudre que des problèmes de très faible taille.

- **nature de l'approche :** c'est déjà très bien d'avoir une approche analytique qui décrive les performances d'un système. Mais c'est souvent au prix d'hypothèses fortes. Aussi, est-il souhaitable de posséder un outil qui permette d'évaluer les stratégies sur une même base. Il suffit ici d'avoir un outil qui décrive les performances d'un système donné dans un contexte le plus réaliste possible. La simulation colle parfaitement à cette description. Quant à un outil recherchant une stratégie optimale, c'est là beaucoup demander. Il en existe cependant un dans un certain contexte pour la famille des tactiques sans relocalisations.

6.9.2. Sans relocalisation

6.9.2.1 Problèmes de recouvrement

Les premiers outils ont été développés dans un contexte déterministe. Des demandes sont issues des noeuds d'un réseau (avec des pourcentages variables) et pour y répondre on dispose de serveurs qui sont supposés toujours disponibles. C'est donc bien la ressource la plus près de la demande qui y répond. En voici quelques exemples dans la littérature :

Dans [Toregas, 1972] on impose une borne supérieure sur la distance entre tout point de demande et son serveur le plus près. On cherche le nombre mimimal de serveurs en solutionnant un problème de recouvrement. Il suffit de résoudre le programme linéaire sans contraintes d'intégrité et d'ajouter une coupe au besoin. Dans tous les cas testés l'auteur rapporte que cela a été suffisant. [Bach, 1984] utilise cette méthodologie pour la ville de Dortmund en Allemagne de l'ouest.

[Daskin, 1981] va un peu plus loin dans la même veine. Pour tenir compte du fait que le serveur le plus près n'est pas nécessairement disponible, il modifie l'objectif pour favoriser les couvertures multiples. On parle alors de "hierarchical objective set covering" puisque l'on a en fait deux objectifs :

1) minimiser le nombre de ressources nécessaires pour couvrir chaque noeud à l'intérieur d'un temps T fixé a priori;
2) étant donné ce nombre minimal trouvé en 1), maximiser la somme sur tous les points de demande, des ressources couvrant un point de demande.

[Eaton, 1985] fait état d'une vaste étude menée pour la ville d'Austin au Texas. Le mandat général était de déterminer le nombre et la localisation des ressources de manière à assurer une qualité de service à tous les segments de la population (Blancs, Noirs, personnes agées, appels urgents, appels non-urgents, etc.) et ce à un coût raisonnable. Après avoir passé en revue diverses approches possibles (problème de recouvrement, simulation, modèle stochastique, etc.) il a finalement opté pour le "maximal covering location model". Dans ce modèle on pondère chacun des objectifs visés (huit dans le cas d'Austin) et on maximise cette pondération pour un nombre donné de véhicules. Ce modèle a avant tout été retenu pour sa simplicité. Contrairement au problème de recouvrement on n'exige pas que tous les noeuds soient couverts. Cela peut réduire sensiblement le nombre de véhicules requis.

6.9.2.2 Modèles stochastiques

En pratique, chaque véhicule n'est pas toujours disponible. Les modèles stochastiques visent à tenir compte explicitement de cet aspect et ainsi combler la lacune principale des modèles présentés ci-haut. La nature aléatoire des systèmes d'urgence est multiforme. Il y a d'abord la demande qui est distribuée aléatoirement dans le temps mais également dans l'espace. Il y a ensuite le service dont la durée est inconnue à l'avance et finalement l'état du système (le statut, occupé ou non, de chacun des serveurs).

a) le modèle hypercube

Le modèle stochastique le plus important pour les systèmes d'urgence est certes l'hypercube [Larson, 1974], [Larson, 1981]. Dans ce modèle, la région à analyser est modélisée par un ensemble de sous-régions chacune représentée par un noeud. À chaque noeud est associée la fraction des appels émanant de cette sous-région. On connaît également le temps de déplacement entre toutes paires de noeuds. Pour répondre aux appels on dispose de véhicules qui, lorsqu'ils sont libres, peuvent être positionnés sur un ou des noeuds (dans ce dernier cas on modélise une patrouille en spécifiant la probabilité que la ressource soit en chacun de ces noeuds). Les localisations de véhicule sont fixées et aucune relocalisation n'est permise.

On suppose ensuite que les demandes surviennent dans la région selon un processus de Poisson de paramètre constant dans le temps, chaque noeud agissant lui-même comme un générateur Poisson indépendant. Pour chaque appel une seule ressource est requise. Le choix du véhicule à affecter dépend d'une règle de préférence qui peut varier d'un noeud à l'autre et qui est fixée a priori. Lorsque tous les serveurs sont

occupés, on peut soit décider de refuser l'appel, utiliser une ressource spéciale supposée toujours disponible ou constituer une file d'attente gérée par une politique indépendante du lieu de l'appel. Le temps de service de chaque véhicule est supposé distribué exponentiellement et indépendant de l'identité du serveur, la localisation de l'appel et le passé du système.

L'état du système est représenté par un ensemble ordonné ayant autant de composantes que de véhicules. Chaque composante symbolise l'état d'une ressource : occupée ou libre. Dans un système sans file d'attente, il y a donc $2^{\text{nombre de véhicules}}$ états possibles. L'ensemble des états forme un hypercube unitaire dans $R^{\text{nombre de véhicules}}$, d'où le nom du modèle. Les transitions entre les états sont facilement indentifiées. On obtient un processus Markovien à nombre discret d'états et à temps continu. Larson propose ensuite une méthode pour calculer les probabilités des états à l'équilibre de manière exacte. On peut ainsi résoudre des problèmes avec un nombre d'ambulances allant jusqu'à 10 ou 12. À partir des probabilités à l'équilibre on peut calculer une foule de statistiques sur les performances du système. Le modèle hypercube est un modèle descriptif qui à partir de localisations données en entrée nous renseigne sur la qualité de la solution, ce qui est particulièrement utile lorsqu'on s'est donné plusieurs objectifs. Il n'est pas question ici de trouver des emplacements optimaux par rapport à un certain objectif. Les performances générées peuvent toutefois implicitement suggérer des améliorations.

b) Ajouts à l'hypercube

Plusieurs ajouts ont été faits au travail initial de Larson. Dans [Chelst, 1979] on développe la mécanique pour calculer la fonction de distribution du temps de transport (à ne pas confondre avec le temps de

réponse qui lui inclut une attente possible en file d'attente). Cela devient utile lorsqu'on veut étudier, par exemple, la qualité du service dans les cas d'arrêt cardiaque où le nombre de vies sauvées et le temps de transport sont en relation non-linéaire.

Avec le modèle hypercube on pouvait résoudre des problèmes comportant jusqu'à douze ambulances environ. Plusieurs systèmes utilisent un bien plus grand nombre de véhicules à la fois. À Montréal, par exemple, en 1986 on en utilisait jusqu'à 79. Dans [Larson, 1975] on présente une procédure approximative pour calculer les performances du modèle hypercube. Plutôt que de résoudre 2^N équations linéaires pour trouver les probabilités à l'équilibre, on résout N équations non-linéaires par un processus itératif pour trouver la charge de travail de chaque véhicule. Essentiellement les mêmes statistiques que dans le modèle exact peuvent ensuite être déduites. Pour les cas où il y a pu comparer avec le modèle exact, l'auteur rapporte des erreurs de moins de 2%.

Dans [Brandau, 1983] on retrouve une application de l'hypercube au service ambulancier de Boston. En plus de résultats on y présente trois améliorations au modèle :

- pour les cas où il n'y a pas de file d'attente, on permet dans la méthode approximative (contrairement à l'original [Larson, 1975]) des temps de service moyens dépendant du serveur;
- calibration du temps moyen de service (TMS) : la valeur du TMS d'un serveur dépend d'une constante (qui représente le temps moyen de service qui n'est pas du transport, i.e. qui va de l'arrivée sur les lieux jusqu'au retour en disponibilité) et du temps moyen de transport du serveur (TMTS). Or le TMTS calculé par le modèle utilise une valeur initiale du TMS et donc une valeur initiale implicite du TMTS. Entre

ces deux valeurs du TMTS (initiale et calculée par le modèle) il risque d'y avoir incompatibilité. On a donc mis au point un processus itératif qui se poursuit tant que les deux valeurs du TMTS ne sont pas suffisamment près;

- algorithme de barrière : dans le cas où on n'a pas de réseau explicite, l'algorithme pour le calcul de la table du temps de déplacement entre toutes paires de centroïdes est basé sur la position des centroïdes, une métrique pour la distance (Manhattan) et une vitesse moyenne. On tient compte explicitement de barrières (rivières, montagnes) polygonales pour ajuster ces temps.

Pour un excellent tour d'horizon de la problématique sur la localisation dans un contexte hypercube, nous suggérons [Larson, 1978b].

L'hypercube est un outil descriptif. [Jarvis, 1975] l'a inséré dans un processus itératif en vue de trouver un ensemble de localisations optimal. L'objectif est de minimiser le temps moyen de transport. La procédure ne garantit pas la convergence vers un optimum global et de fait ne l'atteint pas toujours. Chaque itération consiste à d'une part fixer la règle d'affectation des véhicules aux appels (la règle du "plus près" est utilisée) et trouver les meilleurs emplacements dans ce contexte; puis d'autre part fixer les emplacements et recalculer la règle d'affectation.

c) Districts indépendants d'un seul serveur

[Berman, 1985c] considère un réseau avec deux serveurs seulement et dont la localisation est donnée a priori, chacun travaillant de manière indépendante (sans coopération) selon un système M/G/1 sur son territoire (avec file d'attente possible traitée FIFO). On veut déterminer la meilleure partition de la région en deux zones de façon à minimiser le temps moyen de réponse (temps d'attente et temps de transport) pour un

au hasard. Des résultats exacts sont obtenus pour des valeurs extrêmes du taux d'appel (très petit et très élevé) et des heuristiques présentées pour un taux quelconque d'appel.

On tient compte explicitement de la dépendance du temps de service par rapport à la localisation du serveur. La méthode peut se généraliser pour plus de deux serveurs. La solution optimale peut s'éloigner sensiblement de la règle "du plus près" (le territoire d'un serveur et l'ensemble de noeuds dont il est le plus près) et elle est très sensible au taux global d'appel.

[Berman, 1984] considère le problème général de localisation- allocation en contexte de congestion où des files d'attente peuvent se former. On veut simultanément trouver la meilleure combinaison d'emplacements des ressources (localisations) et de partition du territoire à raison d'un territoire par ressource (allocation) de manière à minimiser le temps de réponse. La demande provient des noeuds du réseau et les ressources peuvent être localisées n'importe où sur le réseau (noeuds ou arcs). L'heuristique présentée solutionne alternativement un problème de localisation d'un véhicule pour chaque district [Berman, 1985a] et un problème d'allocation en généralisant les résultats contenus dans [Berman, 1985c].

d) "Dispatching"

[Berman, 1982b] étudie le contexte où tous les véhicules libres sont fixés à l'exception d'un seul qui est en mouvement vers une destination connue. On peut imaginer une ambulance qui revient de l'hôpital vers sa caserne ou qui est en relocalisation. En supposant le trajet du véhicule en mouvement connue à l'avance (il se déplace à une vitesse constante connue) et en supposant une règle d'affectation des ressources qui envoie toujours la plus près de l'appel, le modèle détermine :

- pour chaque véhicule, la probabilité qu'il réponde au prochain appel;
- l'espérance du temps de réponse au prochain appel.

On suppose implicitement qu'il n'y a aucun retour en disponibilité de ressources qui seraient en service au moment où débute l'analyse. Dans le cas contraire il faudrait tenir compte de cette ressource qui pourrait être la plus près d'un appel.

[Berman, 1983a] généralise ces résultats au cas de plusieurs serveurs en mouvement. Etant donné que l'appel provient d'un certain noeud, on construit la table des intervalles de temps où un serveur donné est le plus près de l'appel.

Dans presque tous les modèles de répartition on suppose que les décisions quant au nombre de serveurs et à leur localisation sont déjà prises. Dans ce cadre, [Carter, 1972] a étudié le cas de deux véhicules en coopération. Étant donné une distribution spatiale quelconque des taux d'appels (il n'y a pas de réseau sous-jacent ici), on cherche le découpage du territoire en deux zones de première intervention pour chaque véhicule qui minimise le temps moyen de réponse. Les auteurs montrent que ce n'est pas toujours la règle "la ressource libre la plus près prend l'appel" qui est la meilleure. Dans le modèle hypercube cette fois (avec des serveurs différentiables) mais avec plus de deux serveurs, [Jarvis, 1981] propose un algorithme pour trouver une politique optimale d'affectation des serveurs aux appels (un serveur par appel). À cause de la taille du problème, on a recours à une caractérisation de la solution optimale, qui est inscrite dans un processus d'approximations successives pour les processus de renouvellement Markovien. Les résultats, pour des problèmes de petite taille (jusqu'à dix serveurs et 50 noeuds), indiquent des gains minimes (de l'ordre de 2%) par rapport à la règle usuelle "le serveur le plus près est affecté à l'appel". Dans le

même contexte, [Berman,] ne suppose pas que tous les serveurs sont soumis à une seule et même autorité. Il peut y en avoir plusieurs (des municipalités, par exemple) avec chacune ses préoccupations. La question de la coopération entre ces autorités peut être étudiée. La décomposition de Benders est suggérée pour résoudre le problème.

e) Affectation de plusieurs véhicules par appel

Dans tous les modèles stochastiques vus jusqu'ici on faisait l'hypothèse qu'un seul serveur était nécessaire pour répondre à un appel. La réalité peut être différente. Elle l'est sûrement pour le service des incendies mais elle peut également l'être pour les urgences médicales. Nous abordons maintenant quelques articles qui se sont penchés sur cet aspect.

[Green, 1980] aborde le cas où chaque appel requiert un nombre aléatoire de serveurs (tous identiques, les appels n'ont pas de répartition spatiale mais que temporelle). Le service de chacun d'eux débute au même instant mais tous ne terminent pas au même moment. On fait de plus l'hypothèse que le service ne peut débuter que lorsque le nombre de ressources nécessaires est disponible, ce qui est peu réaliste pour une urgence médicale.

Toujours dans le même contexte, [Green, 1981] étudie diverses politiques de gestion de la file d'attente. Il apparaît que la règle FIFO n'est pas toujours la plus efficace par rapport à divers critères. On tire profit du fait que des serveurs peuvent être libres lorsqu'il y a des appels en file d'attente car chacun attend un nombre approprié de serveurs libres.

En reprenant pas à pas la démarche de [Larson, 1974], [Chelst, 1981] introduit dans le modèle hypercube la possibilité qu'une ou deux ressources soient nécessaires pour répondre à un appel. On ne considère que le cas sans file d'attente.

Dans [Green, 1984a] on étudie un modèle de file d'attente avec plusieurs niveaux de priorité et plusieurs serveurs pour un même appel. Le nombre de serveurs requis est aléatoire et dépend du type d'appel et du nombre de serveurs disponibles. On s'intéresse à l'espérance mathématique du temps d'attente pour les appels des différentes classes d'appel. Il n'est pas question de l'aspect spatial du problème et aucun résultat sur les temps de réponse (transport + délai) n'est présenté.

La presque totalité des modèles de file d'attente pour système d'urgence suppose qu'un seul serveur est nécessaire pour répondre à un appel. Dans ces variantes des modèles M/M/N on peut être tenté de modifier certains paramètres pour simuler l'effet qu'un nombre aléatoire de serveurs soit requis. [Green, 1984b] montre que ni l'ajustement du taux d'appel, ni celui du temps de service et ni celui du nombre de serveurs ne donne de résultats satisfaisants.

[Daskin, 1984a] développe un modèle de la distribution du temps de transport pour le premier de plusieurs véhicules requis. Le temps de parcours des arcs est distribué normalement et la covariance des temps de parcours des arcs successifs d'un même chemin est supposé nulle (ce qui est peu réaliste). Les temps de parcours des chemins des divers véhicules requis eux sont corrélés. Pour cette raison, dans certains cas les chemins des véhicules affectés à un appel ne devront pas tous être les chemins (en moyenne) les plus courts.

f) Systèmes à plusieurs priorités avec tactique de type plancher

Par définition, dans un système à plusieurs priorités un ensemble d'ambulances doit répondre à des appels de diverses priorités. Les appels de plus grande priorité sont évidemment servis les premiers, peu importe le moment où ils sont survenus. Pour mettre d'avantage l'emphase sur les appels des priorités supérieures, on peut mettre en place une tactique de type plancher. Concrètement, à chaque classe de priorité i

est associé à un nombre c_i tel que si un appel de priorité i survient et que le nombre total de serveurs disponibles est inférieur à c_i l'appel sera placé en file d'attente. Un tel système est en place chez Urgences santé. On a trois classes de priorité avec c_1=0, c_2=25, c_3=25.

[Taylor, 1980] étudie un tel système avec deux classes de priorité. On y calcule la probabilité que n serveurs soient occupés, l'espérance du temps d'attente pour la classe de faible priorité et (dans le cas où il n'y a pas de file d'attente pour la classe de première priorité) la distribution de ce temps d'attente. Ce n'est pas un modèle spatial de file d'attente comme l'hypercube et il n'est jamais question du temps de réponse. [Schaack, 1985] généralise au cas d'un nombre quelconque de classes de priorité.

g) Les systèmes AVL

Depuis quelques années, des systèmes AVL ("automatic vehicle location") ont fait leur apparition sur le marché. Cet outil permet au répartiteur de connaître le véhicule qui est vraiment le plus près de l'appel. C'est surtout utile lorsque les véhicules n'occupent pas un noeud fixé mais plusieurs avec une certaine probabilité (comme les patrouilles de la police). Dans le contexte de l'hypercube, [Larson, 1978a] présente un modèle analytique de cette situation et des résultats qui indiquent les plus grandes améliorations pour les taux d'occupation les plus faibles. La solution proposée résout exactement les équations à l'équilibre. Aucune méthode approximative permettant d'attaquer des problèmes de plus grande taille n'est proposée.

Dans un certain contexte, [Berman, 1982a] montre qu'il existe une solution optimale où toutes les localisations sont sur des noeuds du réseau (et donc aucune sur un arc). Le contexte est le même que celui décrit pour l'hypercube sauf que le temps de service est de distribution

générale (pas nécessairement exponentielle) mais indépendante du temps de transport (et donc indépendant de la paire (serveur, lieu de l'appel)). On suppose également que le temps de transport sur un arc est distribué au pro rata de la distance. Ce résultat indique que l'hypercube n'est pas handicapé en ne considérant que l'ensemble des noeuds comme localisations potentielles.

Pour le même problème mais avec un seul véhicule on n'a pas besoin de supposer le temps de service indépendant du temps de transport. Dans ce contexte [Berman, 1985a] propose des algorithmes pour trouver l'emplacement qui minimise le temps moyen de réponse (transport + délai en file d'attente). Cet emplacement peut être sur un arc du réseau.

h) Réseaux stochastiques

La plupart des modèles présentés jusqu'ici supposait déterministe le temps de parcours d'un arc du réseau. En pratique les conditions changeantes de la circulation font que ce temps de parcours est lui aussi une variable aléatoire. Nous énumérons ici quelques articles abordant ce sujet.

[Berman, 1982c] aborde une variante du problème du p-médian. On définit l'état d'un réseau par la donnée des temps de parcours de chacun des arcs. Les transitions entre les états sont supposées Markoviennes. En tout temps l'état du réseau est connu et par conséquent les temps de parcours aussi. À chaque changement d'état, on peut relocaliser (instantanément) les serveurs avec un certain coût. Le but est de localiser et relocaliser p serveurs toujours disponibles où la fonction objective est une pondération du temps de réponse et des coûts de relocalisation. On montre que sous certaines conditions une solution optimale sur les noeuds du réseau existe. Une heuristique pour le cas d'un seul serveur est présentée. [Berman, 1983c] généralise pour un nombre quelconque de serveurs. De même, alors que dans [Berman, 1982c] les serveurs doivent

toujours occuper les mêmes positions lorsque le réseau est dans un état donné, dans [Berman, 1985b] on permet que la nouvelle configuration dépende de la précédente.

Toujours en supposant les serveurs disponibles en permanence, [Weaver, 1983] considère un nombre fini d'états où le temps de parcours des arcs et la demande aux noeuds varient.

Enfin, [Daskin, 1984b] présente un modèle multi-objectif où il traite simultanément des problèmes de localisation, répartition et des chemins empruntés. Le réseau est stochastique au sens où les temps de parcours sont non-déterministes. Aucune relocalisation n'est permise et les serveurs sont disponibles en permanence. Chaque appel peut nécessiter un ou deux serveurs.

i) Divers

[Larson, 1982] a étendu le modèle hypercube au contexte policier où chaque serveur peut être dans l'un de trois états :

- occupé sur un appel;
- occupé à une autre activité prise à son initiative;
- libre et patrouillant.

On a donc un système de 3^N états (N=nombre de serveurs) et Larson le résout exactement ou approximativement comme pour l'hypercube.

Au moment d'un retour en disponibilité ou d'un repositionnement, un véhicule est en mouvement. Au cours du trajet un appel peut survenir et ce véhicule peut contribuer à améliorer le temps de réponse. Il est pertinent de se demander quel est le meilleur chemin à emprunter pour atteindre une destination fixe. [Berman, 1983b] étudie le sujet pour le cas d'un seul véhicule en mouvement et où ce véhicule est en plus le seul sur le réseau.

6.9.2.3 Approche hybride

Dans le problème de couverture maximale on tente avec M véhicules de maximiser la demande couverte à l'intérieur d'un certain temps de réponse. La principale lacune de ce modèle est de supposer les serveurs toujours disponibles. [Daskin, 1983] tente d'y ajouter un contenu stochastique en supposant connue la probabilité p qu'un serveur soit occupé (la même pour tous les serveurs). On fait également l'hypothèse que la probabilité qu'un serveur i soit occupé est indépendante de celle qu'un serveur j le soit (ce qui est peu réaliste mais plus plausible que le modèle déterministe). On formule un problème de maximisation de l'espérance de la couverture et une heuristique est proposée pour construire progressivement une solution pour chaque valeur de p, en débutant avec p=1,0. [Daskin, 1982] applique ce modèle sur deux réseaux de faible taille.

[Weaver, 1985] considère le problème du p-médian où chaque noeud n'est pas nécessairement servi par le véhicule le plus près. Pour chaque noeud i on suppose connus des coefficients P_{ik} donnant le pourcentage des appels provenant de ce noeud et servis par le $k^{ième}$ serveur le plus près du noeud ($\sum_{k=1}^{p} P_{ik}=1$). On démontre qu'une solution optimale existe où tous les serveurs sont sur des noeuds du réseau.

6.9.3 Avec relocalisations

Dans [Wesolowsky, 1975], en tout temps G serveurs sont localisés parmi M candidats des N points du réseau. On désire élaborer une stratégie de localisation et de relocalisation optimales étant donné des changements connus dans le niveau et la répartition de la demande sur un horizon de k périodes. À chaque période, des relocalisations peuvent être effectuées à un certain coût et les serveurs sont supposés toujours disponibles.

[Kolesar, 1973] considère le contexte du service des incendies où on désire par des relocalisations s'assurer d'une couverture minimale. Cela signifie que pour chaque point de demande, un camion-pompe et un camion-échelle soient suffisamment près. Il y a relocalisation dès que cette couverture minimale n'est pas assurée. L'approche comporte quatre sous-problèmes solutionnés séquentiellement :

- identifier si la couverture minimale est violée et qu'une relocalisation est nécessaire;
- pour les points dont la couverture minimale n'est pas respectée, identifier le nombre minimal de casernes qui en assure la couverture;
- identifier les ressources relocalisées, en respectant la couverture minimale et minimisant le temps moyen de réponse;
- déterminer l'affectation des ressources relocalisées aux casernes qui minimise la distance totale parcourue.

On suppose implicitement qu'aucun incendie ne surviendra durant la relocalisation.

En utilisant la méthodologie des processus de décision Markovien, [Berman, 1981a] examine analytiquement dans le contexte du modèle hypercube la relocalisation des véhicules. Pour chaque état du réseau on permet la relocalisation d'au plus un véhicule. Il s'agit de trouver la politique optimale qui minimise l'espérance des coûts à long terme. On y fait les hypothèses suivantes. S'il survient un appel, on affecte toujours la ressource la plus près, en incluant une éventuelle ressource en relocalisation dont on suppose connue la position exacte en tout temps. Il n'y a pas de file d'attente lorsque tous les serveurs sont occupés. À la fin d'un service, le véhicule se dirige toujours vers le poste d'attente libre le

plus près (excluant un poste vers lequel se dirigerait un véhicule en repositionnement). On suppose enfin que durant la relocalisation au plus un événement (début ou fin de service) peut se produire.

[Berman, 1981b] étend ces résultats au cas où les serveurs ne sont plus identiques, au sens suivant : chaque serveur est maintenant rattaché à un poste d'attente spécifique duquel il peut être relocalisé au besoin vers un ensemble de points environnant. À la fin d'un service, chaque ressource retourne à son poste d'attente. On examine deux cas : i) le répartiteur ignore la position d'une ressource libre en mouvement (au cours d'un repositionnement) et ne peut l'utiliser lorsqu'un appel survient, ii) le répartiteur connaît parfaitement la position de tous les véhicules disponibles.

6.9.4 Nos réalisations

En ce qui a trait aux stratégies opérationnelles, nous nous étions fixés comme objectifs de :

- couvrir le plus large éventail de contextes;
- identifier la meilleure stratégie possible pour chacun d'eux;
- développer un outil d'évaluation des performances d'une stratégie qui reflète le mieux son comportement réel.

La revue de la littérature nous a montré l'étendue des travaux dans le domaine. Dans bien des cas des modèles analytiques sont présentés et solutionnés. Ils fournissent alors plusieurs renseignements utiles sur les performances du système étudié. Ils font cependant des hypothèses non seulement variables d'un système à l'autre mais surtout très fortes. En général on suppose que les demandes forment un processus de Poisson (ce qui est réaliste) et les temps de service sont distribués exponentiellement (ce qui est moins réaliste mais n'est pas très dérangeant). Mais là où on

s'éloigne d'avantage de la réalité c'est lorsqu'on suppose atteint un état d'équilibre et que le taux global d'appel est constant. Ces hypothèses sont évidemment faites parce que c'est le seul moyen de solutionner les modèles mathématiques élaborés. Ceci ne signifie pas qu'ils sont sans intérêts, bien au contraire, mais il faut être prudent dans leur interprétation. Et cette prudence est d'autant plus de mise dans le contexte des services d'urgence que des vies humaines sont en jeu. La justesse des conclusions s'impose. C'est la raison pour laquelle une large partie de nos efforts au chapitre des stratégies opérationnelles a été consacrée à la réalisation d'un outil permettant de reproduire le plus fidèlement possible les conditions prévalant dans la réalité. Cet outil [Choquette, 1987a], [Choquette, 1987b] permet entre autre de voir dans quelle mesure les performances annoncées par les modèles analytiques se vérifient dans la "réalité".

6.9.4.1 Sans relocalisation

Aucune des approches déterministes ne nous est apparue intéressante en soi. Elles sont d'abord fortement handicapées par l'hypothèse que le serveur le plus près est toujours disponible, et des modèles plus complets sont présentement disponibles. Ces modèles déterministes ne sont toutefois pas complètement inutiles. En particulier, lorsque le taux d'occupation des serveurs est faible on n'est pas loin de l'hypothèse que le serveur le plus près est toujours disponible et ces modèles déterministes peuvent fournir d'excellentes solutions, voire une solution initiale à des algorithmes plus sophistiqués. Ils peuvent également servir pour attaquer certains sous-problèmes. Ce fut le cas pour Urgences santé où on retrouve un certain nombre de postes d'attente où la couverture doit être permanente et où les relocalisations sont permises pour s'en assurer. On peut ainsi définir certains objectifs par rapport à ces postes puis faire l'hypothèse qu'ils

sont toujours occupés ce qui rend le problème déterministe. Chez Urgences santé l'objectif choisi fut de minimiser le nombre de ces postes d'attente de manière à couvrir au moins 95% des demandes à l'intérieur de 5 minutes et 100% des demandes à l'intérieur de 10 minutes.

Nous avons donc implanté quelques algorithmes issus de modèles déterministes. Il y a d'abord un algorithme de p-médian, inspiré des suggestions de [Larson, 1981]. Il y a également un algorithme de p-centre inspiré de [Handler, 1979] auquel nous avons ajouté une procédure récursive particulièrement efficace. Il y a enfin quelques variantes de ces deux problèmes. Ainsi dans le problème du p-centre on fixe le nombre de véhicules et on désire minimiser la distance maximale d'un noeud au serveur le plus près. On peut aussi s'intéresser à la situation réciproque où on fixe une distance maximale et on recherche la configuration respectant cette distance maximale en utilisant le moins de serveurs possible. Dans le même contexte on peut vouloir que seulement un certain pourcentage p_1 des demandes respecte une distance maximale d_1 et que 100% respecte une plus grande distance maximale d_2. Ce fut le cas chez Urgences santé où p_1=0,95, d_1=5 minutes et d_2=10 minutes.

En ce qui a trait au contexte le plus contraint où les districts sont indépendants, nous n'avons rien fait. Nos efforts ont plutôt porté sur le cas des districts dépendants où les serveurs sont en coopération. Dans ce domaine le modèle hypercube de Larson nous est apparu de loin le plus intéressant. Nous avons donc implanté le modèle hypercube solutionné de manière exact. Ceci permet, pour un réseau de la taille de celui utilisé à Montréal (484 noeuds), d'analyser une flotte de dix ambulances et moins. Nous avons aussi implanté l'algorithme qui permet d'obtenir une solution approximative du modèle hypercube. Dans ce cas nous avons étudié des problèmes ayant jusqu'à soixante véhicules et on peut envisager de plus

grandes tailles encore. Puisque ces deux modèles ne sont que descriptifs, nous leur avons adjoint le processus iteratif de Jarvis qui recherche les localisations optimales. Les premiers résultats ont toutefois montré que, contrairement aux prétentions de ce dernier, la présence d'optimum locaux peut causer des problèmes. Nous avons tenté de corriger la situation avec un nouvel algorithme [Trudeau, 1987a] qui s'avère significativement plus efficace que celui de Jarvis.

Puis nous avons adapté ces travaux à un contexte légèrement différent de celui de l'hypercube. Plutôt que de disposer de N véhicules et de pouvoir les localiser où que ce soit sur le réseau on impose que ces N véhicules doivent se retrouver en au plus C casernes ($C < N$) [Trudeau, 1987b]. Nous simulons ainsi une situation de contrainte économique où on ne peut se permettre de bâtir N casernes. Ce contexte ressemble à celui d'Urgences santé, bien qu'eux permettent des relocalisations. Ils ont en effet présentement un nombre de postes d'attente inférieur (de beaucoup) au nombre de véhicules en service à quelque moment du jour que ce soit.

Enfin, nous avons aussi implanté l'algorithme de Daskin pour le modèle d'espérance de couverture maximale. Cette approche devrait produire des solutions inférieures à celles de Larson en raison des hypothèses faites. En voulant vérifier si c'était bien le cas, nous nous sommes butés à des temps d'exécution astronomiques, à cause de la taille de nos problèmes. Nous n'avons pas poursuivi dans cette veine.

Ces outils que nous avons développés, nous permettent de traiter très adéquatement le domaine des stratégies sans relocalisations. Il reste à faire des ajouts de la nature de ceux décrits ci-haut en 6.9.2.2 e), f), g), h). À noter qu'au niveau du "dispatching" nous avons toujours utilisé la règle "le serveur disponible le plus près répond à l'appel", nous appu-

yant sur les travaux de [Jarvis, 1981].

6.9.4.2. Avec relocalisation

Le fait de permettre aux véhicules d'être relocalisés engendre un problème très complexe. En effet, alors que les véhicules sont à effectuer une relocalisation et qu'ils sont en mouvement il peut survenir des appels et des retours en disponibilité. Cela vient modifier la composition des véhicules disponibles de sorte qu'à la fin de la relocalisation on se retrouve dans un état que l'on ne peut connaître a priori.

Pour encore mieux saisir la problématique, supposons au contraire qu'aucun événement ne survient durant la relocalisation. Cela revient à faire l'hypothèse que la relocalisation a lieu instantanément, les véhicules se déplaçant à une vitesse infinie. Alors si notre seul objectif est de minimiser le temps moyen de réponse, et si k véhicules sont disponibles, la meilleure configuration pour faire face au prochain appel est celle du k-médian. Lorsque le taux d'appel est très faible et que les distances de relocalisation sont courtes, il peut être raisonnable de supposer qu'aucun événement ne se produise durant une relocalisation.

Berman a supposé qu'au plus un événement pouvait survenir durant une relocalisation. Et même en faisant une hypothèse aussi forte (du moins pour le contexte de Montréal) il n'a étudié que le cas où on relocalisait un seul véhicule à la fois et ce pour des problèmes de petite taille. Ceci est une indication de la complexité du problème.

Nous sommes plutôt intéressés à analyser la relocalisation de plusieurs véhicules à la fois, pour des problèmes de grande taille et où plusieurs événements risquent de se produire durant la relocalisation. Dans l'état actuel de la recherche aucune méthode analytique ne peut aborder ce problème. Mais disposant d'un outil de simulation on peut imaginer et évaluer diverses stratégies. Même si à cet égard nos travaux ne sont pas

complétés, des résultats préliminaires ont été obtenus. Nous avons d'abord cherché à mesurer théoriquement le gain maximal que pouvait procurer la relocalisation instantanée décrite plus tôt. Puis par simulation nous avons mesuré le comportement de cette stratégie lorsque le nombre d'événements au cours de la relocalisation n'était pas nul. Enfin quelques variations ont été introduites, notamment sur la fréquence des relocalisations. Tout cela a été fait en comparaison avec les meilleures stratégies sans relocalisations, avec nombre limité de casernes ou non [Trudeau, 1987c].

En conclusion, beaucoup reste à faire au niveau stratégique. Il faut d'abord pousser plus loin le développement d'heuristiques pour le contexte avec relocalisation. D'autre part, avec ou sans relocalisation, il reste toute une autre avenue à explorer où le taux d'appel varie. Dans ce contexte le système n'a pas le temps d'atteindre l'état d'équilibre. Comment chaque période où le taux d'appel peut être supposé constant est-elle influencée par la précédente? Dans tous ces cas la simulation sera un outil privilégié.

RÉFÉRENCES

[Allaire, 1986] : Allaire, B., Baril, J.M., Cloutier, P., Vanier, P., "Department of Ambulance Services (Toronto), compte rendu visite à Toronto 12-13-14 mars 1986".

[Bach, 1984] : Bach, L., "Locational Optimization for Health Facilities - A Case Study for Emergency Dispatch in Dortmund". 3rd int. conf. on System Science in Health Care, Springer Verlag, 1036-1039 (1984).

[Berman, 1981a] : Berman, O., "Dynamic Repositioning of Indistinguishable Service Units on Transportation Networks". Transportation Science, vol. 15, no 2, 115-136 (1981).

[Berman, 1981b] : Berman, O., "Repositioning of Distinguishable Urban Service Units on Networks". Comput & Ops Res., vol. 8, 105-118 (1981).

[Berman, 1981c] : Berman, O., Larson, R.C., Odoni, A.R., "Developments in Network Location With Mobile and Congested Facilities". EJOR, vol. 6, 104-116 (1981).

[Berman, 1982a] : Berman, O., Larson, R.C., "The Median Problem with Congestion". Comput & Ops Res., vol. 9, no 2, 119-126 (1982).

[Berman, 1982b] : Berman, O., "Dispatching Service Units on Networks Using Perfect Real Time Location Information". INFOR, vol. 20, no 4, 357-369 (1982).

[Berman, 1982c] : Berman, O., Odoni, A.R., "Locating Mobile Servers on a Network with Markovian Properties". Networks, vol. 12, 73-86 (1982).

[Berman, 1983a] : Berman, O., Rahmana, M.R., "A Procedure for Dispatching Moving Mobile Servers". Networks, vol. 13, 121-141, (1983).

[Berman, 1983b] : Berman, O., Rahmana, M.R. "Optimal Path of a Single Service Unit on a Network to a 'Nonemergency' Destination". Transportation Science, vol. 17, no 2, 218-232, (1983).

[Berman, 1983c] : Berman, O., Leblanc, B., "Location-relocation of N Mobile Facilities on a Stochastic Network". Univ. of Calgary, octobre 1983.

[Berman, 1984] : Berman, O., Mandowsky, R.R., "Location-allocation on Congested Networks". WP-08-84, Univ. of Calgary (1984).

[Berman, 1985a] : Berman, O., Larson, R.C., Chiu, S.S., "Optimal Server Location on a Network Operating as an M/G/I Queue". Operations Research, vol. 33, no 4, 746-771, juillet-août 1985.

[Berman, 1985b] : Berman, O., Rahmana, M.R., "Optimal Location-relocation Decisions on Stochastic Networks". Transportation Science, vol. 19, no 3, 203-221 (1985).

[Berman, 1985c] : Berman, O., Larson, R.C., "Optimal 2-facility Network Districting in the Presence of Queuing". Transportation Science, vol. 19, no 3, 261-277 (1985).

[Bertolazzi, 1976] : Bertolazzi, P., Bianco, L., Ricciardelli, S., "A Method for Determining the Optimal Districting in Urban Emergency Services". R. 76-07, Universita di Roma, Istituto di Automatica (1976).

[Blais, 1978] : Blais, J.Y., "Manuel d'utilisation de HASTUS-macro", Centre de recherche sur les transports, Université de Montréal, publication no 93, février 1978.

[Box, 1976] : Box, G.E.P., Jenkins, G.M., "Time Series Analysis: Forecasting and Control", Holden-Day (1976).

[Brandau, 1983] : Brandau, M.L., Larson, R.C., "Extending and Applying the Hypercube Queuing Model to Deploy Ambulances in Boston". December 1983.

[Carter, 1972] : Carter, G.M., Chaiken, J.M., Ignall, E., "Response Areas for Two Emergency Units". Operations Research, vol. 20, no 3, 571-594 (1972).

[Chelst, 1979] : Chelst, K., Jarvis, J.P., "Estimating the Probability Distribution of Travel Times for Urban Emergency Service Systems". Operations Research, vol. 27, no 1, 199-204 (1979).

[Chelst, 1981] : Chelst, K.R., Barlach, Z., "Multiple Unit Dispatches in Emergency Services: Models to Estimate System Performance". Management Science, vol. 27, no 12, 1390-1409 (1981).

[Choquette, 1987] : Choquette, J., Trudeau, P., Ferland, J., Rousseau, J.M., "Un modèle de simulation pour un système d'Urgences santé : guide de l'usager". Centre de recherche sur les transports, Université de Montréal, (à paraître).

[Dagenais, 1980] : Dagenais, M.G., Gaudry, M.J.I., Liem, T.C., "Multiple Regression Analysis with Box-Cox Transformations and Nonspherical Residual Errors: A Transportation Application", publication no 166, Centre de recherche sur les transports, Université de Montréal, avril 1980.

[Daskin, 1981] : Daskin, M.S., Stern, E.H., "A Hierarchical Objective Set Covering Model for Emergency Medical Service Vehicle Deployment". Transportation Science, vol. 15, no 2, 137-152 (1981).

[Daskin, 1982] : Daskin, M.S., "Application of an Expected Covering Model to Emergency Medical Service System Design". Decision Sciences, vol. 13, no 3, 416-439 (1982).

[Daskin, 1983] : Daskin, M.S., "A Maximum Expected Covering Location Model: Formulation, Properties and Heuristic Solution". Transporation Sciences, vol. 17, no 1, 48-70 (1983).

[Daskin, 1984a] : Daskin, M.S., Haghani, A., "Multiple Vehicle Routing and Dispatching to an Emergency Scene". Environment and Planning A, vol. 16, 1349-1359 (1984).

[Daskin, 1984b] : Daskin, M.S., "Location, Dispatching and Routing Models for Emergency Services with Stochastic Travel Times". Transportation Center, Northwestern University, juin 1984.

[Eaton, 1985] : Eaton, D.J., Daskin, M.S., Simmons, D., Bullock, B., Jansma, G., "Determining Emergency Medical Service Vehicle Deployment in Austin, Texas". Interfaces, vol. 15, no 1, 96-108 (1985).

[Green, 1980] : Green, L., "A Queueing System in Which Customers Require a Random Number of Servers". Ops. Res., vol. 28, no 6, 1335-1346 (1980).

[Green, 1981] : Green, L., "Comparing Operating Characteristics of Queues in Which Customers Require a Random Number of Servers". Management Science, vol. 27, no 1, 65-74 (1981).

[Green, 1984a] : Green, L., "A Multiple Dispatch Queueing Model of Police Patrol Operations". Management Science, vol. 30, no 6, 653-664 (1984).

[Green, 1984b] : Green, L., Kolesar, P., "A Comparison of the Multiple Dispatch and M/M/C Priority Queueing Models of Police Patrol". Management Science, vol. 30, no 6, 665-670 (1984).

[Groom, 1975] : Groom, K.N., "The Estimation of Emergency Ambulance Cover", Royal Institute of Public Administration, National Health Service Operational Research Group. Rapport no 75/16, décembre 1975.

[Handler, 1979] : Handler, G.Y., Mirchandani, P.B., "Location on Networks, Theory and Algorithms", MIT Press, Cambridge, Mass. (1979).

[Heyman, 1982] : Heyman, D.P., Sobel, M.J., "Stochastic Models in Operations Research Volume 1, Stochastic Processes and Operating Characteristics". McGraw-Hill, N.Y. (1982).

[Heyman, 1984] : Heyman, D.P., Sobel, M.J., "Stochastic Models in Operations Research Volume 2, Stochastic Optimization". McGraw-Hill, N.Y. (1984).

[Holloway, 1976a] : Holloway, K.E., Raitt, R.A., Groom, K.N., "Ambulances in Wales. The Experience of Eight Operational Research Studies". Royal Institute of Public Administration, National Health Service Operational Research Group, rapport no 76/16, mai 1976.

[Holloway, 1976b] : Holloway, K.E., "Dyfed Ambulances, Nonemergency Services". Royal Institute of Public Administration, National Health Service Operational Research Group, rapport no 76/25, novembre 1976.

[Howard, 1960] : Howard, R.A., "Dynamic Programming and Markov Processes". M.I.T. Press, Cambridge, Mass. (1960).

[Jarvis, 1975] : Jarvis, J.P., "Optimization in Stochastic Service with Distinguishable Servers". TR-19-75, Innovative Resource Planning Project in Urban Public Safety Systems, Operations Research Center, M.I.T., juin 1975.

[Jarvis, 1981] : Jarvis, J.P., "Optimal Assignments in a Markovian Queuing System". Comput & Ops. Res., vol. 8, 17-23 (1981).

[Kolesar, 1974] : Kolesar, P., Walker, W.E., "An Algorithm for the Dynamic Relocation of Fire Companies". Operations Research, vol. 22, no 2, 249-274 (1974).

[Larson, 1972] : Larson, R.L., Stevenson, K.A., "On Insensitivities in Urban Redistricting and Facility Location". Operations Research, vol. 20, 595-612 (1972).

[Larson, 1974] : Larson, R.L., "A Hypercube Queueing Model for Facility Location and Redistricting in Urban Emergency Services". Comput & Operations Research, vol. 1, 67-95 (1974).

[Larson, 1975] : Larson, R.L., "Approximating the Performance of Urban Emergency Service Systems". Operations Research, vol. 23, no 5, 845-868 (1975).

[Larson, 1978a] : Larson, R.L., Franck, E.A., "Evaluating Dispatching Consequences of Automatic Vehicle Location in Emergency Services". Comput. & Operations Research, vol. 5, 11-30 (1978).

[Larson, 1978b] : Larson, R.L., "Structural System Models for Locational Decisions: An Example Using the Hypercube Queuing Model". OR'78. North Holland Publishing Company (1979), K.B. Haley ed., 1054-1091.

[Larson, 1981] : Larson, R.L., Odoni, A.R., "Urban Operations Research", Prentice-Hall Inc., Englewood Cliffs, N.J. (1981).

[Larson, 1982] : Larson, R.L., McKnew, M.A., "Police Patrol - Initiated Activities Within a Systems Queueing Model". Management Science, vol. 28, no 2, 759-774 (1982).

[Liem, 1983] : Liem, T.C., Dagenais, M.G., Gaudry, M.J.I., "DSDM Project L-1.1: A Program for Box-Cox Transformations in Regression Models with Heteroskedastic and Autoregressive Residuals". Centre de recherche sur les transports, publication no 301, Université de Montréal, juin 1983, révisé en avril 1984.

[Mayer, 1981] : Mayer, J.D., "A Method for the Geographical Evaluation of Emergency Medical Service Performance". AJPH, vol. 71, no 8, 841-844 (1981).

[Ontario, 1970] : "Ambulance Act and Regulation 14", gouvernement de l'Ontario.

[Orcon, 1974] : "Ambulance Service Performance Standards and Measurement. The Final Report". Orcon Services. Le 12 juillet 1974.

[Québec, 1983] : Arrêté ministériel du 30 juin 1983, Gazette officiel du Québec, 115e année, no 39, 4013.

[Raitt, 1976a] : Raitt, R.A., Kempner, V.A., MacMahon, K., "Emergency Ambulance Cover in East Sussex". Royal Institute of Public Administration, National Health Service Operational Research Group, rapport no 76/2, janvier 1976.

[Raitt, 1976b] : Raitt, R.A., Holloway, K.E., "Emergency Ambulance Cover in Dyfed". Royal Institute of Public Administration, National Health Service Operational Research Group, rapport no 76/15, juin 1976.

[Raitt, 1977] : Raitt, R.A., "Progress in Ambulance Service Planning", Royal Institute of Public Adminstration, National Health Service Operational Research Group, rapport no 77/2, janvier 1977.

[Ross, 1970] : Ross, M., "Applied Probability Models with Optimization Applications". Holden-Day, San Francisco (1970).

[Rousseau, 1978] : Rousseau, J.M., Dupuis, D., Blais, J.Y., "Affectation des chauffeurs avec le système HASTUS-macro", Centre de recherche sur les transports, Université de Montréal, publication no 113, septembre 1978.

[Schaack, 1985] : Schaack, C., Larson, R.L., "An N Server Cutoff Multipriority Queue". Working Paper OR135-85, M.I.T. (1985).

[Taylor, 1980] : Taylor, I.D.S., Templeton, J.G.C., "Waiting Time in a Multi-server Cutoff-priority Queue, and its Application to an Urban Ambulance Service". Operations Research, vol. 28, no 5, 1168-1188 (1980).

[Teitz, 1968] : Teitz, M.B., Bart, P., "Heuristic Methods for Estimating the Generalized Vertex Median of a Weighted Graph". Operations Research, vol. 16, 955-961 (1968).

[Toregas, 1971] : Toregas, C., Swain, R., ReVelle, C., Bergman, L., "The Location of Emergency Service Facilities". Operations Research, vol. 19, 1363-1373 (1971).

[Toronto, 1986] : "Metropolitan Toronto Department of Ambulances Services".

[Trudeau, 1987a] : Trudeau, P., Hodgson, J., "Localisation optimale de véhicules d'urgence". Centre de recherche sur les transports, Université de Montréal, (à paraître).

[Trudeau, 1987b] : Trudeau, P., "Localisations optimales de véhicules d'urgence lorsque le nombre de postes d'attente est limité". Centre de recherche sur les transports, Université de Montréal, (à paraître).

[Trudeau, 1987c] : Trudeau, P., Choquette, J., "Simulation de diverses stratégies de déploiement de véhicules d'urgence". Centre de recherche sur les transports, Université de Montréal, (à paraître).

[Trudeau, 1987d] : Trudeau, P., Choquette, J., Hodgson, J., Ferland, J., Rousseau, J.M., "Un modèle de simulation pour un système d'Urgences santé : la problématique". Centre de recherche sur les transports, Université de Montréal, (à paraître).

[U.K., 1976] : "Notes on Visit of U.K. Delegation to Moscow and Kiev", UK/USSR/Health Cooperation Agreement: EMS and Traumatology, septembre/octobre 1976.

[U.S., 1974] : "United States Statutes at Large, 1973 and Reorganization Plans and Proclamations", volume 87, United States Government Printing Office, Washington (1974).

[Wesolowsky, 1975] : Wesolowsky, G.O., Truscott, W.G., "The Multiperiod Location-allocation Problem with Relocation of Facilities". Management Science, vol. 22, no 1, 57-65 (1975).

[Weaver, 1983] : Weaver, J.R., Church, R.L., "Computational Procedures for Location Problems on Stochastic Networks". Transportation Science, vol. 17, no 2, 168-180 (1983).

[Weaver, 1985] : Weaver, J.R., Church, R.L., "A Median Location Model with Nonclosest Facility Service". Transportation Science, vol. 19, no 1, 58-84 (1985).